Natural Language Analysis

How to analyze tons of text in minutes
An (incomplete) technical overview

Or: what in the world do I do with all this text?

William Wood Harter - Luminoso Technologies - wharter@luminoso.com

About Me

Chapman Alumni – Computer Science

Currently at Luminoso since 2018

Software Development and Sales – Hughes, Sun Microsystems, GameWorld.com, FileNet, IBM, Luminoso Technologies

Chapman Adjunct Professor - 1998-2012

HUGHES
HUGHES AIRCRAFT COMPANY

Sun microsystems

GameWorld Technologies

CHAPMAN UNIVERSITY

FILENET

IBM

LUMINOSO

The problem

I have 300, 1k, 10k, 100k, 1M, 10M

{ Product reviews
Open ended survey questions
Trouble tickets
Transcriptions }

do they have any value and what do I do?

Analyzing Text

Text Cleanup

Word counting - Word clouds

Tagging parties

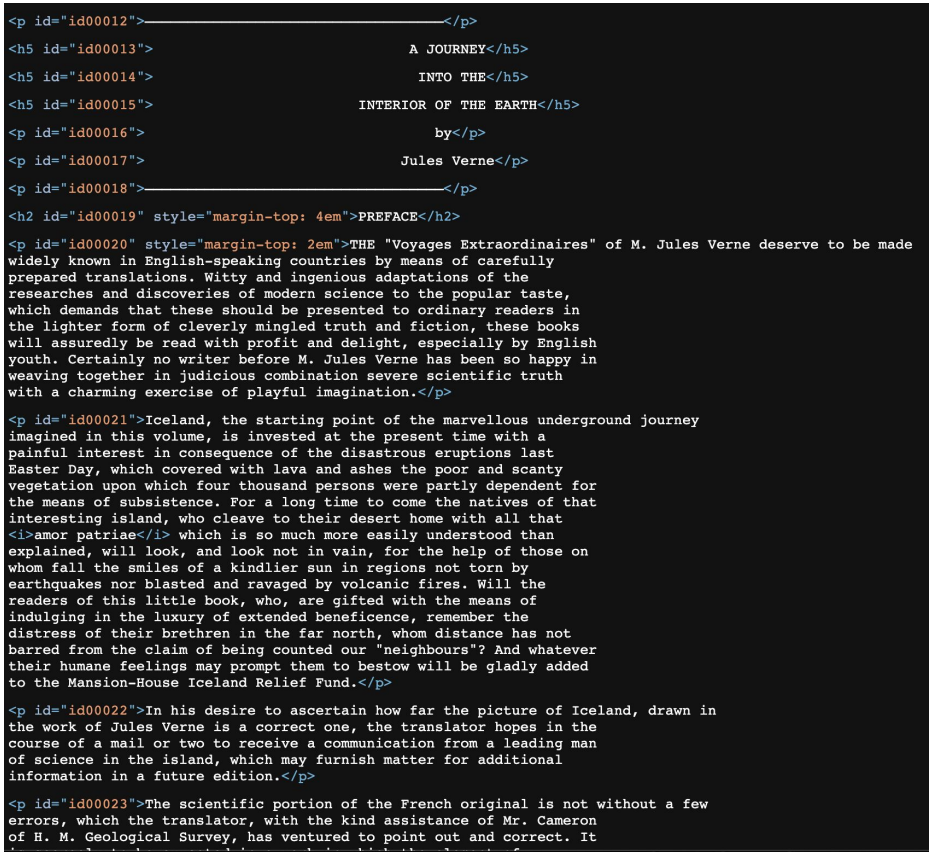
Entity extractors

Vectorization

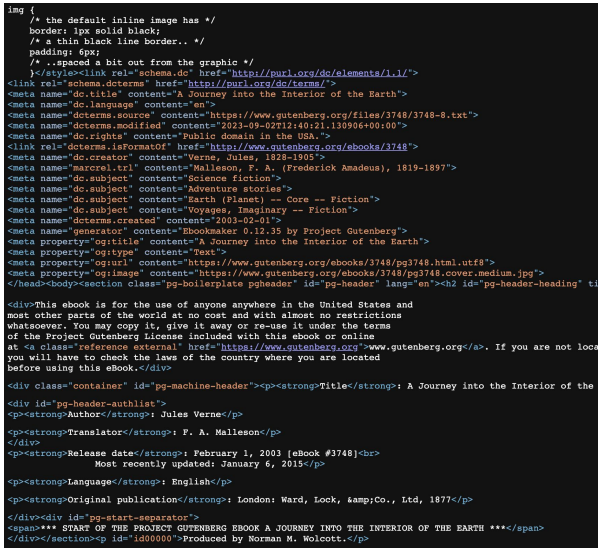
cNN/rNN Convolutional/Recurrent Neural Networks

LLMs

Symbolic AI



```
<p id="id00012">————————————————————————————————————</p>

<h5 id="id00013">                                A JOURNEY</h5>

<h5 id="id00014">                                 INTO THE</h5>

<h5 id="id00015">                            INTERIOR OF THE EARTH</h5>

<p id="id00016">                                    by</p>

<p id="id00017">                                 Jules Verne</p>

<p id="id00018">————————————————————————————————————</p>

<h2 id="id00019" style="margin-top: 4em">PREFACE</h2>

<p id="id00020" style="margin-top: 2em">THE "Voyages Extraordinaires" of M. Jules Verne deserve to be made
widely known in English-speaking countries by means of carefully
prepared translations. Witty and ingenious adaptations of the
researches and discoveries of modern science to the popular taste,
which demands that these should be presented to ordinary readers in
the lighter form of cleverly mingled truth and fiction, these books
will assuredly be read with profit and delight, especially by English
youth. Certainly no writer before M. Jules Verne has been so happy in
weaving together in judicious combination severe scientific truth
with a charming exercise of playful imagination.</p>

<p id="id00021">Iceland, the starting point of the marvellous underground journey
imagined in this volume, is invested at the present time with a
painful interest in consequence of the disastrous eruptions last
Easter Day, which covered with lava and ashes the poor and scanty
vegetation upon which four thousand persons were partly dependent for
the means of subsistence. For a long time to come the natives of that
interesting island, who cleave to their desert home with all that
<i>amor patriae</i> which is so much more easily understood than
explained, will look, and look not in vain, for the help of those on
whom fall the smiles of a kindlier sun in regions not torn by
earthquakes nor blasted and ravaged by volcanic fires. Will the
readers of this little book, who, are gifted with the means of
indulging in the luxury of extended beneficence, remember the
distress of their brethren in the far north, whom distance has not
barred from the claim of being counted our "neighbours"? And whatever
their humane feelings may prompt them to bestow will be gladly added
to the Mansion-House Iceland Relief Fund.</p>

<p id="id00022">In his desire to ascertain how far the picture of Iceland, drawn in
the work of Jules Verne is a correct one, the translator hopes in the
course of a mail or two to receive a communication from a leading man
of science in the island, which may furnish matter for additional
information in a future edition.</p>

<p id="id00023">The scientific portion of the French original is not without a few
errors, which the translator, with the kind assistance of Mr. Cameron
of H. M. Geological Survey, has ventured to point out and correct. It
```



```
img {
    /* the default inline image has */
    border: 1px solid black;
    /* a thin black line border.. */
    padding: 6px;
    /* ..spaced a bit out from the graphic */
    }</style><link rel="schema.dc" href="http://purl.org/dc/elements/1.1/">
<link rel="schema.dcterms" href="http://purl.org/dc/terms/">
<meta name="dc.title" content="A Journey into the Interior of the Earth">
<meta name="dc.language" content="en">
<meta name="dcterms.source" content="https://www.gutenberg.org/files/3748/3748-8.txt">
<meta name="dcterms.modified" content="2023-09-02T12:40:21.130906+00:00">
<meta name="dc.rights" content="Public domain in the USA.">
<link rel="dcterms.isFormatOf" href="http://www.gutenberg.org/ebooks/3748">
<meta name="dc.creator" content="Verne, Jules, 1828-1905">
<meta name="marcrel.trl" content="Malleson, F. A. (Frederick Amadeus), 1819-1897">
<meta name="dc.subject" content="Science fiction">
<meta name="dc.subject" content="Adventure stories">
<meta name="dc.subject" content="Earth (Planet) -- Core -- Fiction">
<meta name="dc.subject" content="Voyages, Imaginary -- Fiction">
<meta name="dcterms.created" content="2003-02-01">
<meta name="generator" content="Ebookmaker 0.12.35 by Project Gutenberg">
<meta property="og:title" content="A Journey into the Interior of the Earth">
<meta property="og:type" content="text">
<meta property="og:url" content="https://www.gutenberg.org/ebooks/3748/pg3748.html.utf8">
<meta property="og:image" content="https://www.gutenberg.org/ebooks/3748/pg3748.cover.medium.jpg">
</head><body><section class="pg-boilerplate pgheader" id="pg-header" lang="en"><h2 id="pg-header-heading" ti

<div>This ebook is for the use of anyone anywhere in the United States and
most other parts of the world at no cost and with almost no restrictions
whatsoever. You may copy it, give it away or re-use it under the terms
of the Project Gutenberg License included with this ebook or online
at <a class="reference external" href="https://www.gutenberg.org">www.gutenberg.org</a>. If you are not loca
you will have to check the laws of the country where you are located
before using this eBook.</div>

<div class="container" id="pg-machine-header"><p><strong>Title</strong>: A Journey into the Interior of the

<div id="pg-header-authlist">
<p><strong>Author</strong>: Jules Verne</p>

<p><strong>Translator</strong>: F. A. Malleson</p>
</div>
<p><strong>Release date</strong>: February 1, 2003 [eBook #3748]<br>
                Most recently updated: January 6, 2015</p>

<p><strong>Language</strong>: English</p>

<p><strong>Original publication</strong>: London: Ward, Lock, &amp;amp;Co., Ltd, 1877</p>

</div><div id="pg-start-separator">
<span>*** START OF THE PROJECT GUTENBERG EBOOK A JOURNEY INTO THE INTERIOR OF THE EARTH ***</span>
</div></section><p id="id00000">Produced by Norman M. Wolcott.</p>
```

A JOURNEY INTO THE INTERIOR OF THE EARTH
CHAPTER I.
THE PROFESSOR AND HIS FAMILY
On the 24th of May, 1863, my uncle, Professor Liedenbrock, rushed into his little house, No. 19 Königstrasse, one of the oldest streets in the oldest portion of the city of Hamburg.
Martha must have concluded that she was very much behindhand, for the dinner had only just been put into the oven.

"Well, now," said I to myself, "if that most impatient of men is hungry, what a disturbance he will make!"

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Text	Text	language	Title	Title	Review I	Rating	date_Date	string_Buy	string_Count	string_Manu	string_Produ	string_Retail
2	RV©alitV© virtuelle avec MSFS 2020, parfait,	Virtual realit	fr	RV©alitV© V	Virtual realit	fr	5	8/21/21	Verified Purc	Canada	HP	HP Reverb G	Amazon
3	best resolution and best speakers on the	second	en	So far so goc	So far so goc	en	4	1/17/22		United State	HP	HP Reverb G	Amazon
4	This was a noticeable visual improvement ove	This was a n	en	Working Gre	Working Gre	en	3	10/13/21	Verified Purc	United State	HP	HP Reverb G	Amazon
5	Excelente	Great	pt	?üëç	?üëç	pt	5	6/1/21	Verified Purc	United State	HP	HP Reverb G	Amazon
6	This item is very cheap and return is hell. Ther	This item is v	en	Heavy and ch	Heavy and ch	en	1	6/9/22	Verified Purc	United State	HP	HP Reverb G	Amazon
7	I recently bought my daughter an oculus and w	I recently bo	en	ONLY FOR U	ONLY FOR U	en	4	12/30/21	Verified Purc	United State	HP	HP Reverb G	Amazon
8	I was happy to receive version 2 of the HP Rev	I was happy	en	Nice upgrade	Nice upgrade	en	5	1/12/22	Verified Purc	Canada	HP	HP Reverb G	Amazon
9	I had to return as it was intermittently poweri	I had to retu	en	Sent back	Sent back	en	1	6/19/22	Verified Purc	United Kingd	HP	HP Reverb G	Amazon
10	optimally, and a lot of users seem to get	some	en	Best mid-tie	Best mid-tie	en	5	5/31/21		United State	HP	HP Reverb G	Amazon
11	I am not going to criticize those who left posit	I am not goir	en	Worst VR Sy	Worst VR Sy	en	1	6/17/22	Verified Purc	United State	HP	HP Reverb G	Amazon
12	I bought this a while back to replace my oculu	I bought this	en	Headset is a	Headset is a	en	3	4/5/22		United State	HP	HP Reverb G	Amazon
13	good!!	properly	en	Like new! fa	Like new! fa	en	5	7/8/22		United State	HP	HP Reverb G	Amazon
14	This headset has amazing resolution. I was ab	This headset	en	Perfect for vi	Perfect for vi	en	5	11/1/21	Verified Purc	United State	HP	HP Reverb G	Amazon
15	replaces them under warrenbty. I called HP	on this	en	No warrenty	No warrenty	en	1	11/18/21	Verified Purc	United State	HP	HP Reverb G	Amazon
16	videos. It is very good for MSFS 2020,	awesome	en	Great for VR	Great for VR	en	5	12/20/21		United State	HP	HP Reverb G	Amazon
17	this headset is so mediocre. I have heard	by saying	en	Poor build qu	Poor build qu	en	2	4/14/22	Verified Purc	United State	HP	HP Reverb G	Amazon
18	the controller and it cant be peeled off. Ive	plastic	en	Unimaginabl	Unimaginabl	en	1	2/9/22	Verified Purc	United State	HP	HP Reverb G	Amazon
19	time and spent a considerable amount of	this and a	en	Spectacular	Spectacular	en	3	2/7/21		United State	HP	HP Reverb G	Amazon
20	Really poor experience so far. I received my R	Really poor e	en	Worked for 3	Worked for 3	en	3	3/5/22	Verified Purc	United Kingd	HP	HP Reverb G	Amazon
21	The controller is cheaply made. The vibration i	The controlle	en	Low quality,	Low quality,	en	1	9/4/21	Verified Purc	United State	HP	HP Reverb G	Amazon
22	It's great when it's working but it's super finic	It's great wh	en	Very buggy	Very buggy	en	2	11/20/21	Verified Purc	United State	HP	HP Reverb G	Amazon
23	fantastic , among the best available and the	Pixel	en	Pixel Density	Pixel Density	en	3	6/1/21		United State	HP	HP Reverb G	Amazon
24	VR headset. I pre-ordered mine and waited	good	en	Excellent wh	Excellent wh	en	2	5/19/21		United State	HP	HP Reverb G	Amazon
25	It was terrible box rip apart parts missing with	It was terribl	en	Missing part	Missing part	en	1	11/29/21	Verified Purc	United State	HP	HP Reverb G	Amazon
26	Never got it to work correctly. Lots of issues w	Never got it	en	Good Seller;	Good Seller;	en	3	7/12/21	Verified Purc	United State	HP	HP Reverb G	Amazon
27	Please dear user, do not buy this from	this	en	Wonderful w	Wonderful w	en	2	5/28/21		United State	HP	HP Reverb G	Amazon
28	I love this headsets besides it tiny flaws that r	I love this he	en	Amazing hea	Amazing hea	en	4	6/23/21	Verified Purc	United State	HP	HP Reverb G	Amazon
29	The entire VR industry is a wreck, devices are	The entire VI	en	It's disgustin	It's disgustin	en	1	1/3/22	Verified Purc	United State	HP	HP Reverb G	Amazon
30	You need something better than a GTX 1060 6	You need sor	en	It works as w	It works as w	en	4	11/23/21	Verified Purc	United State	HP	HP Reverb G	Amazon
31	My husband tried this out for his NASCAR raci	My husband	en	RESTOCKING	RESTOCKING	en	1	8/19/21	Verified Purc	United State	HP	HP Reverb G	Amazon
32	necesariamente Hardware y Software,	the Reverb	es	Frustrante p	Frustrating b	es	4	6/12/21	Verified Purc	United State	HP	HP Reverb G	Amazon

Clean text

A Journey into the Interior of the Earth
by Jules Verne

CHAPTER I.
THE PROFESSOR AND HIS FAMILY

On the 24th of May, 1863, my uncle, Professor Liedenbrock, rushed into his little house, No. 19 Königstrasse, one of the oldest streets in the oldest portion of the city of Hamburg.

Martha must have concluded that she was very much behindhand, for the dinner had only just been put into the oven.

"Well, now," said I to myself, "if that most impatient of men is hungry, what a disturbance he will make!"

"M. Liedenbrock so soon!" cried poor Martha in great alarm, half opening the dining-room door.

"Yes, Martha; but very likely the dinner is not half cooked, for it is not two yet. Saint Michael's clock has only just struck half-past one."

"Then why has the master come home so soon?"

"Perhaps he will tell us that himself."

"Here he is, Monsieur Axel; I will run and hide myself while you argue with him."

And Martha retreated in safety into her own dominions.

I was left alone. But how was it possible for a man of my undecided turn of mind to argue successfully with so irascible a person as the Professor? With this persuasion I was hurrying away to my own little retreat upstairs, when the street door creaked upon its hinges; heavy feet made the whole flight of stairs to shake; and the master of the house, passing rapidly through the dining-room, threw himself in haste into his own sanctum.

But on his rapid way he had found time to fling his hazel stick into a corner, his rough broadbrim upon the table, and these few emphatic words at his nephew:

"Axel, follow me!"

I had scarcely had time to move when the Professor was again shouting after me:

"What! not come yet?"

And I rushed into my redoubtable master's study.

Otto Liedenbrock had no mischief in him, I willingly allow that; but unless he very considerably changes as he grows older, at the end he will be a most original character.

He was professor at the Johannæum, and was delivering a series of lectures on mineralogy, in the course of every one of which he broke into a passion once or twice at least. Not at all that he was over-anxious about the improvement of his class, or about the degree of attention with which they listened to him, or the success which might eventually crown his labours. Such little matters of detail never troubled him much. His teaching was as the German philosophy calls it, 'subjective'; it was to benefit himself, not others. He was a learned egotist. He was a well of science, and the pulleys worked uneasily when you wanted to draw anything out of it. In a word, he was a learned miser.

Germany has not a few professors of this sort.

https://www.gutenberg.org/cache/epub/3748/pg3748.html

Tools

BeautifulSoup - Extract text from html
https://beautiful-soup-4.readthedocs.io/en/latest/

Ftfy - Fix unicode issues
https://ftfy.readthedocs.io/en/latest/

spaCy - tokenization
https://spacy.io/

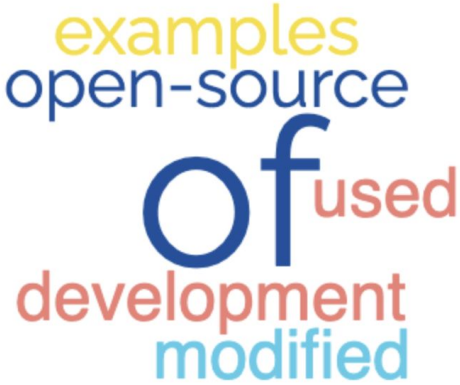
Keras - tokenization
https://keras.io/

Word Counting

```
FILE_NAME = 'file.txt'

wordCounter = {}

with open(FILE_NAME,'r') as fh:
  for line in fh:
    # Replacing punctuation characters. Making the string to lower.
    # The split will spit the line into a list.
    word_list = line.replace(',','').replace('\'','').replace('.','').lower().spli
    for word in word_list:
      # Adding  the word into the wordCounter dictionary.
      if word not in wordCounter:
        wordCounter[word] = 1
      else:
        # if the word is already in the dictionary update its count.
        wordCounter[word] = wordCounter[word] + 1

print('{:15}{:3}'.format('Word','Count'))
print('-' * 18)

# printing the words and its occurrence.
for  (word,occurance)  in wordCounter.items():
  print('{:15}{:3}'.format(word,occurance))
```

#

```
Word           Count
------------------
of              6
examples        2
used            2
development     2
modified        2
open-source     2
```



https://stackoverflow.com/a/42347398/14866475
https://www.wordclouds.com/

Problems with word counting

Problem != Problems

Problme != Problem

FYI != For Your Information

‘A’, ‘the’, ‘why’, ‘I’.... is not a word you want to count

Do you know if ‘problem’ is being used in a positive or negative way?

Tagging Parties

Get a group of co-workers and each person reads 10/100/1000 and tags it for whatever you are looking for (happy/sad/broken button/car battery issue, etc). A lesson in how to frustrate all your co-workers with complete drudgery. Gather the results and graph them.

Entity Extractors

- Detection of
  - People
  - Places (Specifically Addresses)
  - Events
  - Numbers (Phone, CC, SSN)
  - Classification (limited)
  - Emotions (limited)
- Mostly rule based
- Requires Ontology

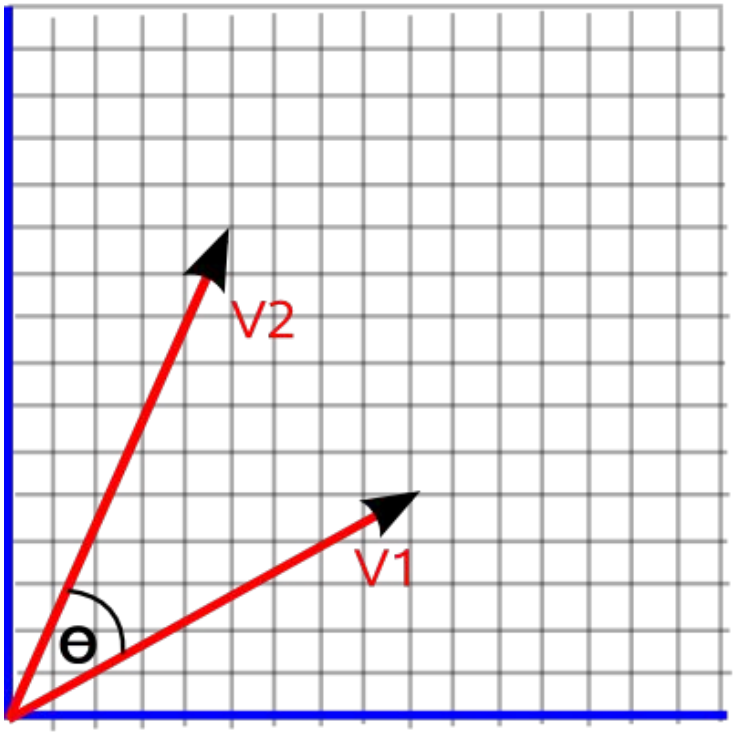
Entity Extraction Example

'Please find my credit card number here: 378282246310005. Thanks for the payment.'

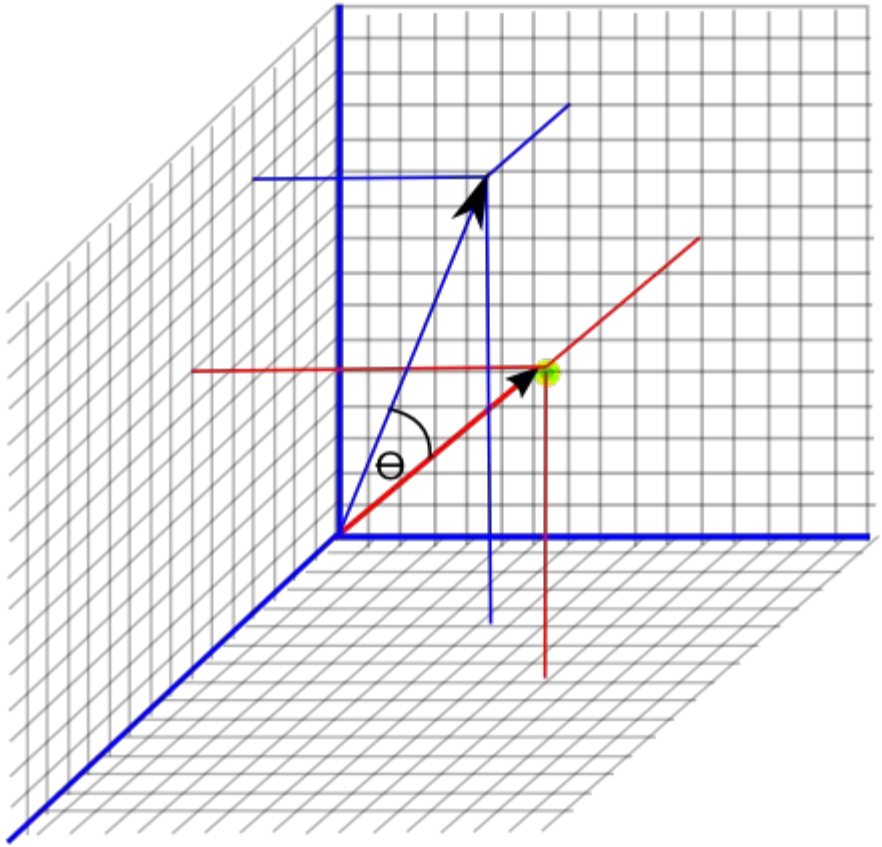

```
{
    "mentions": [
        {
            "span": {
                "begin": 40,
                "end": 55,
                "text": "378282246310005"
            },
            "type": "BankAccountNumber.CreditCardNumber.Amex",
            "producer_id": {
                "name": "RBR mentions",
                "version": "0.0.1"
            },
            "confidence": 0.8,
            "mention_type": "MENTT_UNSET",
            "mention_class": "MENTC_UNSET",
            "role": ""
        }
    ],
    "producer_id": {
        "name": "RBR mentions",
        "version": "0.0.1"
    }
}
```

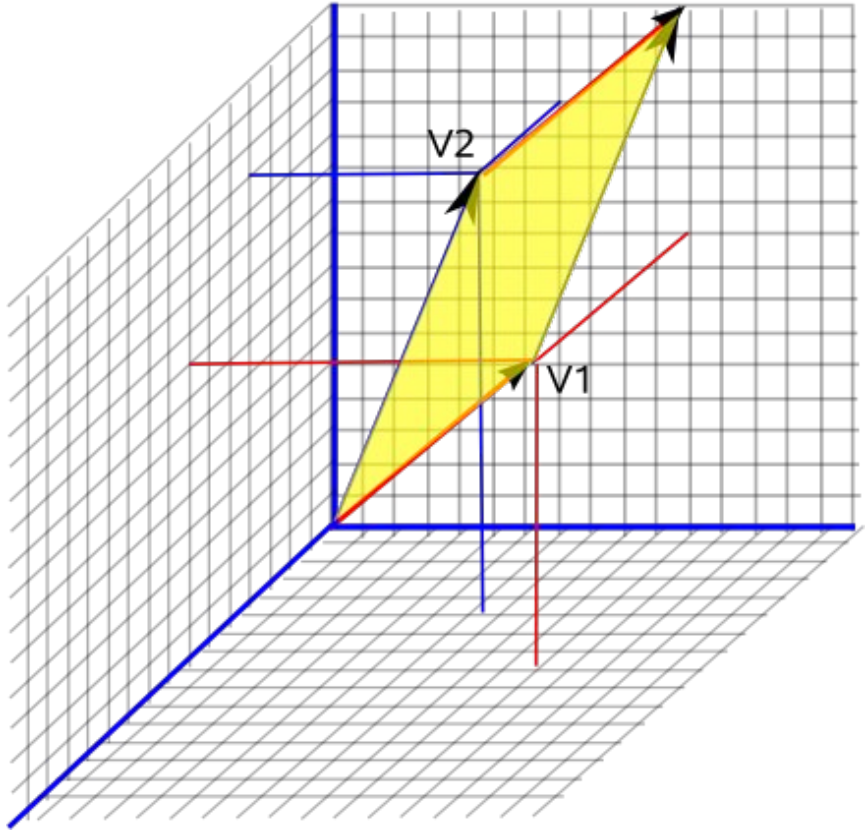

https://www.ibm.com/docs/en/cloud-paks/cp-data/4.7.x?topic=catalog-entity-extraction

Vectorization



https://www.euclideanspace.com/maths/algebra/vectors/vecGeometry/vec2d/vec2dAngle.png





https://www.euclideanspace.com/maths/algebra/vectors/vecGeometry/vec2d/vec2dAngle.png

Matrices

$$\begin{bmatrix} 1 & 9 & -13 \\ 20 & 5 & -6 \end{bmatrix} \quad \begin{bmatrix} 4 & -7 & 5 & 0 \\ -2 & 0 & 11 & 8 \\ 19 & 1 & -3 & 12 \end{bmatrix}$$

https://en.wikipedia.org/wiki/Matrix_(mathematics)

Matrices

$$\begin{bmatrix} 1 & 9 & -13 \\ 20 & 5 & -6 \end{bmatrix}$$

$$\begin{bmatrix} 4 & -7 & 5 & 0 \\ -2 & 0 & 11 & 8 \\ 19 & 1 & -3 & 12 \end{bmatrix}$$

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

https://en.wikipedia.org/wiki/Matrix_(mathematics)

Tokenization

Take the following text:

“There was a warm feeling with the service and I felt like their guest for a special treat.”

https://analyticsindiamag.com/guide-to-text-classification-using-textcnn/

Tokenization - keras example

```
from keras.preprocessing.text import Tokenizer
tokenizer = Tokenizer(num_words=5000)
tokenizer.fit_on_texts(review_train)
Xcnn_train =
tokenizer.texts_to_sequences(review_train)
Xcnn_test =
tokenizer.texts_to_sequences(review_test)
vocab_size = len(tokenizer.word_index) + 1
print(review_train[1])
print(Xcnn_train[1])
```

Output:

```
There was a warm feeling with the service and I felt like their guest for a special treat.
[43, 10, 4, 607, 323, 15, 1, 47, 2, 3, 350, 37, 109, 1908, 12, 4, 279, 1236]
```

https://analyticsindiamag.com/guide-to-text-classification-using-textcnn/

Tokenization: Results

```
There was a warm feeling with the service and I felt like their guest for a special treat.
[43, 10, 4, 607, 323, 15, 1, 47, 2, 3, 350, 37, 109, 1908, 12, 4, 279, 1236]
```

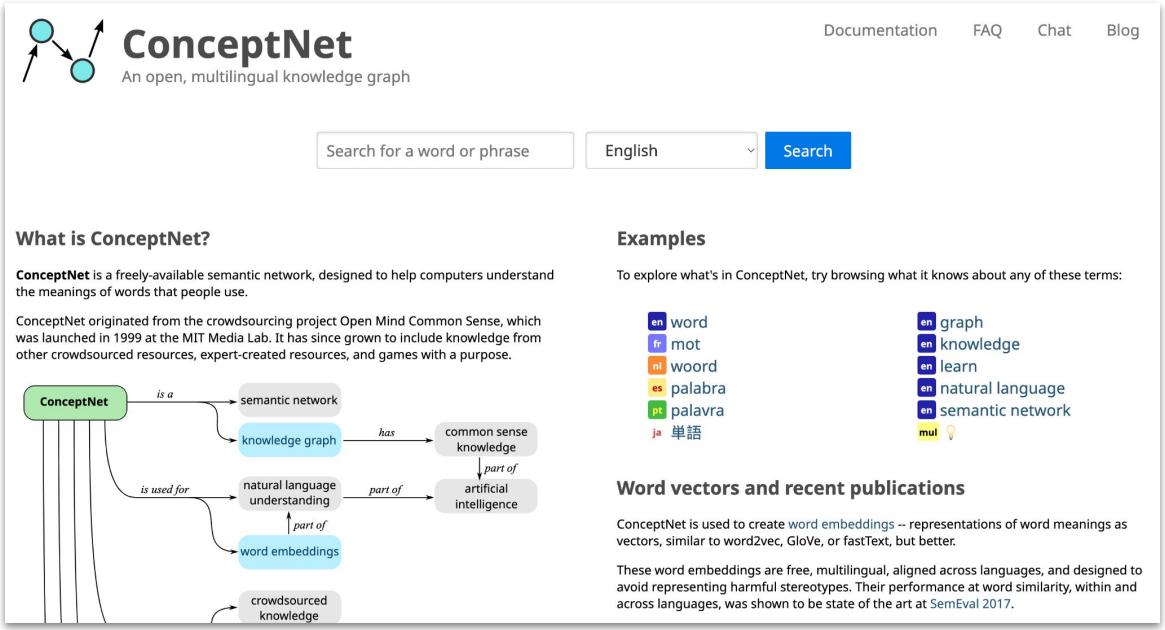
https://analyticsindiamag.com/guide-to-text-classification-using-textcnn/

Tokenization - matrices

```
[ 278  295  212 1907   39  349    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0]
```

https://analyticsindiamag.com/guide-to-text-classification-using-textcnn/

Concepts vs. Documents



word2vec

LUMINOSO

Hugging Face

OpenAI

Vectorization - Word2Vec workflow

Create skip grams of the tokenization result

Iterate over skip grams to generate training data

Build context and label vectors

Build a text vectorization layer - you've create a vocabulary

Build a model with the vocabulary

Get the model weights which has the vectors

https://www.tensorflow.org/text/tutorials/word2vec

Daylight workflow

- CSV or JSON input of text with metadata
- Start build process
  - 20-60 seconds for 2k texts
  - 1-5 minutes for 10k texts
- Read vectors

```
# top concepts
concept_selector = {"type": "top", "limit":50}
top_concepts = client.get('concepts', concept_selector=concept_selector)['result']
top_concepts
```

```
[{'relevance': 301.0,
  'texts': ['dip'],
  'exact_term_ids': ['dip|en'],
  'excluded_term_ids': [],
  'vectors': ['XSh6NeZBWH9Eg9LxA7nFKz_mp_hDF5u8hEBNJAs1_UB9k1B5d_ntBwY_7qCxtBNhBGZChO9Zp_mu_XbAUgEAsB_Q-Vv-Xl_it-eS-e
r_vPAOy_jaAm49U0_ZW92D_RnA2xAoB-AR_th-ldALc9PyBXV_WsAJW_bu9wt9Al8P1BQo_g--65_ihASH_9__4iAzuAwfA0n_8BBGA-lp_9OCns_DABj
3B53-z08VOAvsAf3_rUBU99r2Cz5BjBAxx-uU-A0_10AaR_mCBjj_GL-m-98tAhXD6iALX__7CZlB1hAPB-ZhCVeBue-7A_Ee-gDA5o_pTCKtBh__sW-E
1Af7Bna-5n-u-BDt9LH-vJAHKAdO_ieBTQAZW-Vt_vvAibAHtCYt_5M-qLAhG-OLAg5AEJA3CBuPDwG_dv_blA93_vl_bjAVRBayAWpAo_Ax8BXQ_hVAn
t_oxBGT_pABeeAX-A33A3D-rXAZBAQrAny_RgAT3AsEAfnAvP_o8_xlAyPAWiBsQ_oSCaO-rr-2kAEH_ouA7nBiQ_LK_-LAon_Sd_By-1TAlbANf_sy_G
1AiKBvLBIS_Ff-mTAXhBGL_k__h7AAqA4YBHi_dGAsOAaD9yH-_IAZ7_GQAJ0-R_ARxAElAVIAvpA63BO1A7y_rMAA5_Qs_kD_AUA0C_ZXA5MAZn-hEAG
kAfPAuQACdAgKBCbBOYCBGAHB-R4_gQAK5Bvv_v-_2bBf-_VVAatAw6_QiAX9_3eBUGA0FAL0BN-_zQAwhA6X_8qBO0ASX_9aAbHALoAfJ_44APg_tp_-
oAG-ATU_JQ_n5BMpAc9_ytADaAllA7W_tB_2nAB_AVCAahAGBAJ3__Q_CDAgQAWn_pVAYsAjLAc5AGrAY1Apo_-w__0_mX_8V'],
  'name': 'dip'},
 {'relevance': 282.0,
  'texts': ['laundry'],
```

OpenAI Embeddings

```
curl https://api.openai.com/v1/embeddings \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer $OPENAI_API_KEY" \
  -d '{
    "input": "Your text string goes here",
    "model": "text-embedding-ada-002"
  }'
```


```
{
  "data": [
    {
      "embedding": [
        -0.006929283495992422,
        -0.005336422007530928,
        ...
        -4.547132266452536e-05,
        -0.024047505110502243
      ],
      "index": 0,
      "object": "embedding"
    }
  ],
  "model": "text-embedding-ada-002",
```


https://platform.openai.com/docs/guides/embeddings/what-are-embeddings

Relationships between concepts

Dot Product

```
import numpy as np
concept1 = client_project.get("/concepts")['result'][0]
concept2 = client_project.get("/concepts")['result'][2]
arr1 = np.array([float(v) for v in pack64.unpack64(concept1['vector'])])
arr2 = np.array([float(v) for v in pack64.unpack64(concept2['vector'])])
assoc_score = np.dot(arr1,arr2)
```

Neural Networks

- CNN - Convolutional Neural Networks
  - Mostly for classification
  - Each document vector paired with a label
- RNN - Recurrent Neural Networks
  - Great for time series and text generation
  - Predictive - feed something in, result is what is next

Both use the same kinds of text “vectorization” data.

Building a classification model

```
from keras.models import Sequential
from keras import layers

embedding dim = 200
textcnnmodel = Sequential()
textcnnmodel.add(layers.Embedding(vocab size,
           embedding dim, input length=maxlen))
textcnnmodel.add(layers.Conv1D(128, 5,
            activation='relu'))
textcnnmodel.add(layers.GlobalMaxPooling1D())
textcnnmodel.add(layers.Dense(10, activation='relu'))
textcnnmodel.add(layers.Dense(1, activation='sigmoid'))
textcnnmodel.compile(optimizer='adam',
              loss='binary crossentropy',
              metrics=['accuracy'])
textcnnmodel.summary()
```

https://analyticsindiamag.com/guide-to-text-classification-using-textcnn/

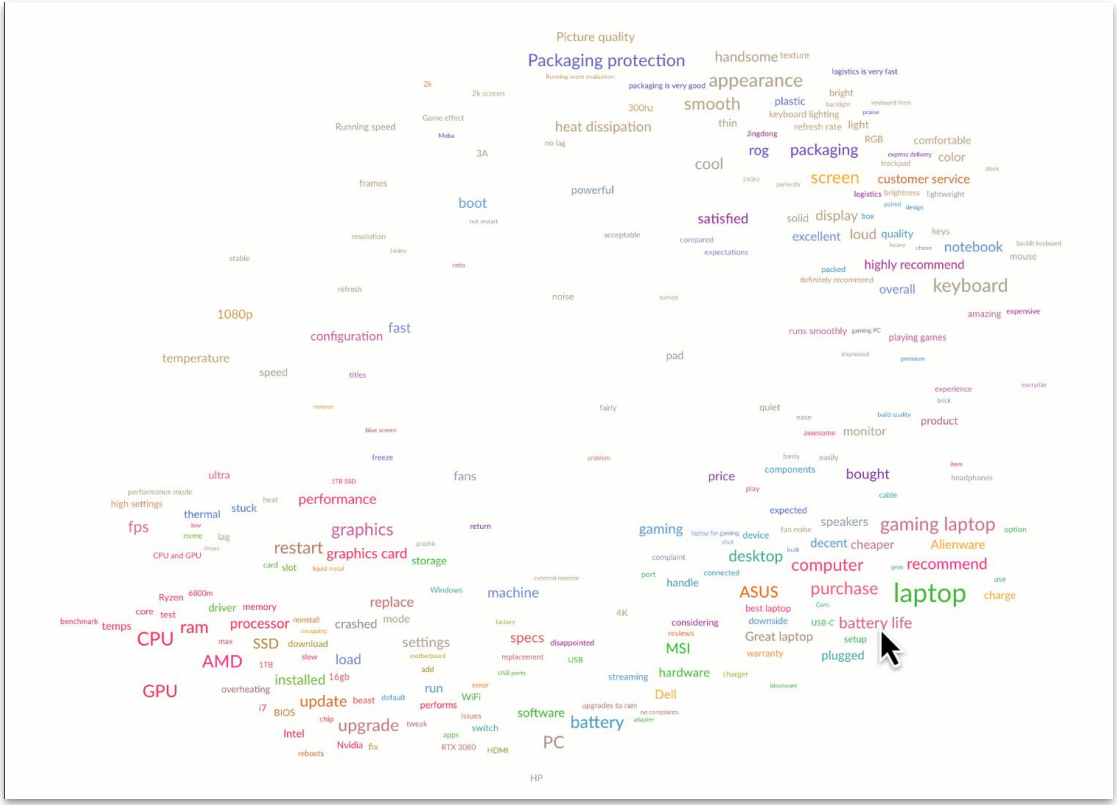
Using a classification model

```
 initial_model = keras.Sequential(
    [
        keras.Input(shape=(250, 250, 3)),
        layers.Conv2D(32, 5, strides=2, activation="relu"),
        layers.Conv2D(32, 3, activation="relu"),
        layers.Conv2D(32, 3, activation="relu"),
    ]
)
feature extractor = keras.Model(
    inputs=initial model.inputs,
    outputs=[layer.output for layer in initial_model.layers],
)

# Call feature extractor on test input.
x = tf.ones((1, 250, 250, 3))
features = feature_extractor(x)
```

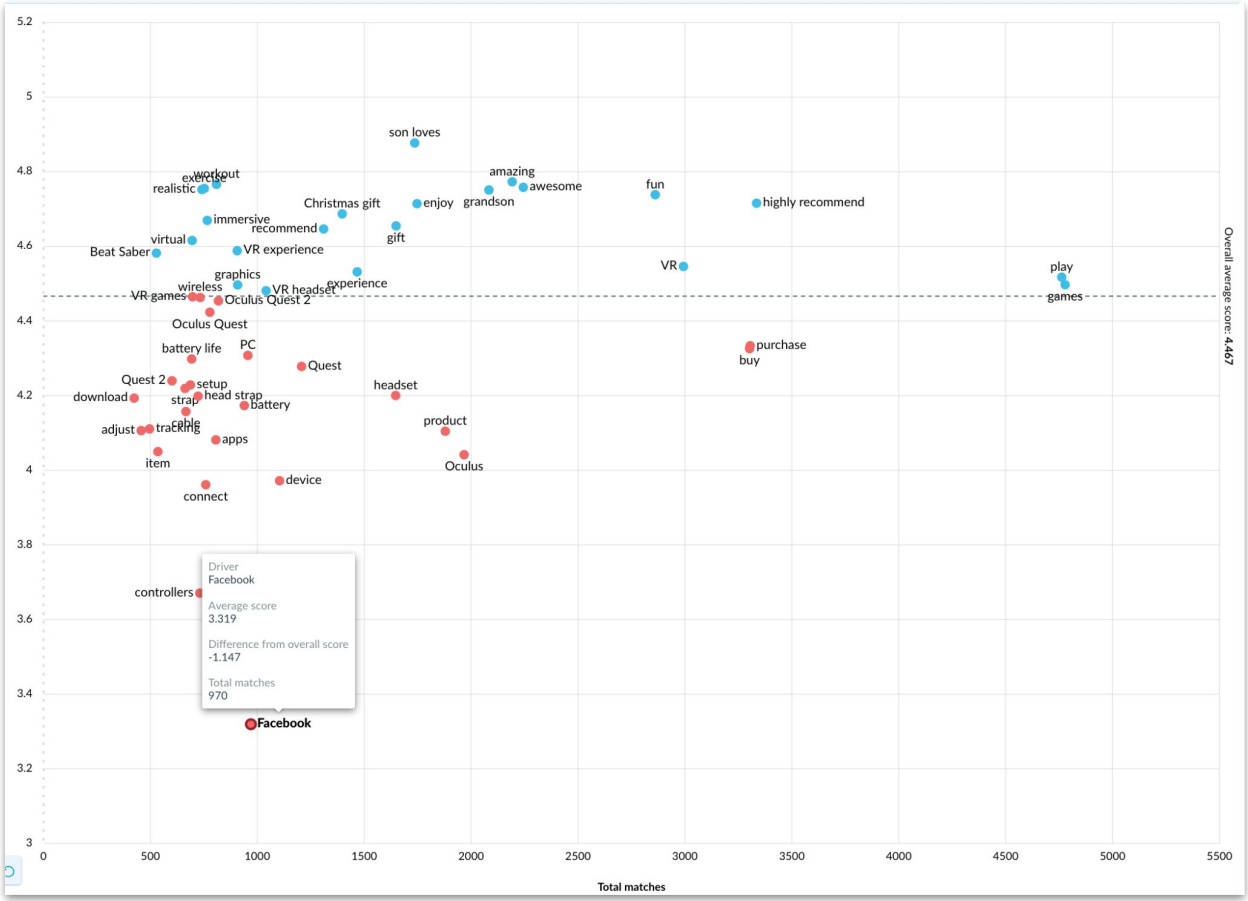
https://keras.io/guides/sequential_model/

Applications

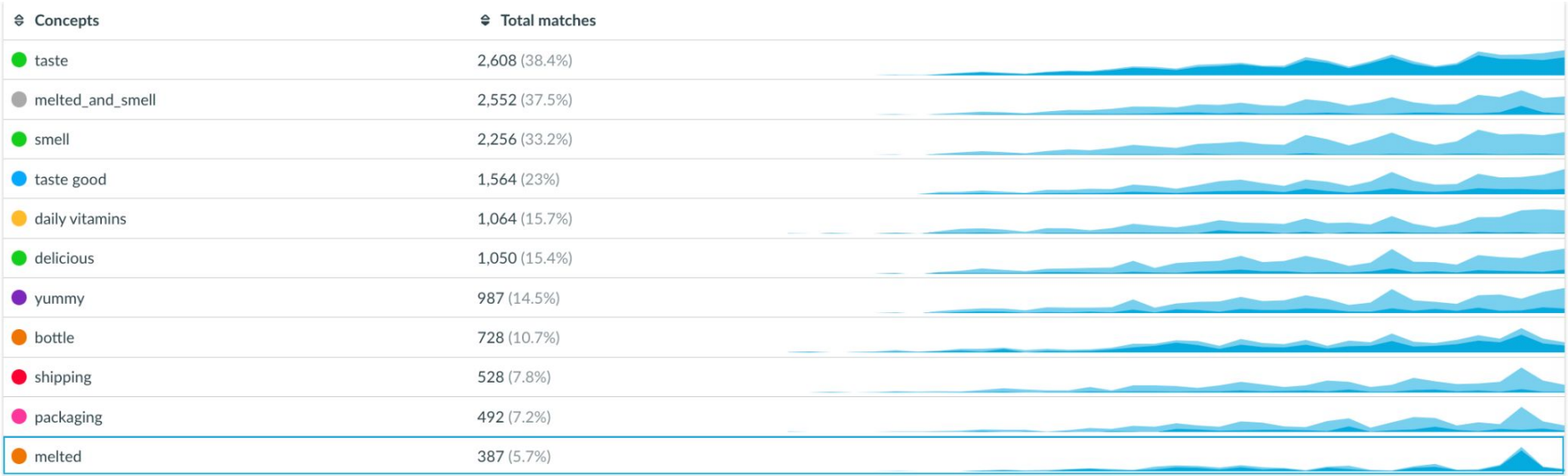
View concept relationships



Measuring Ratings and NPS

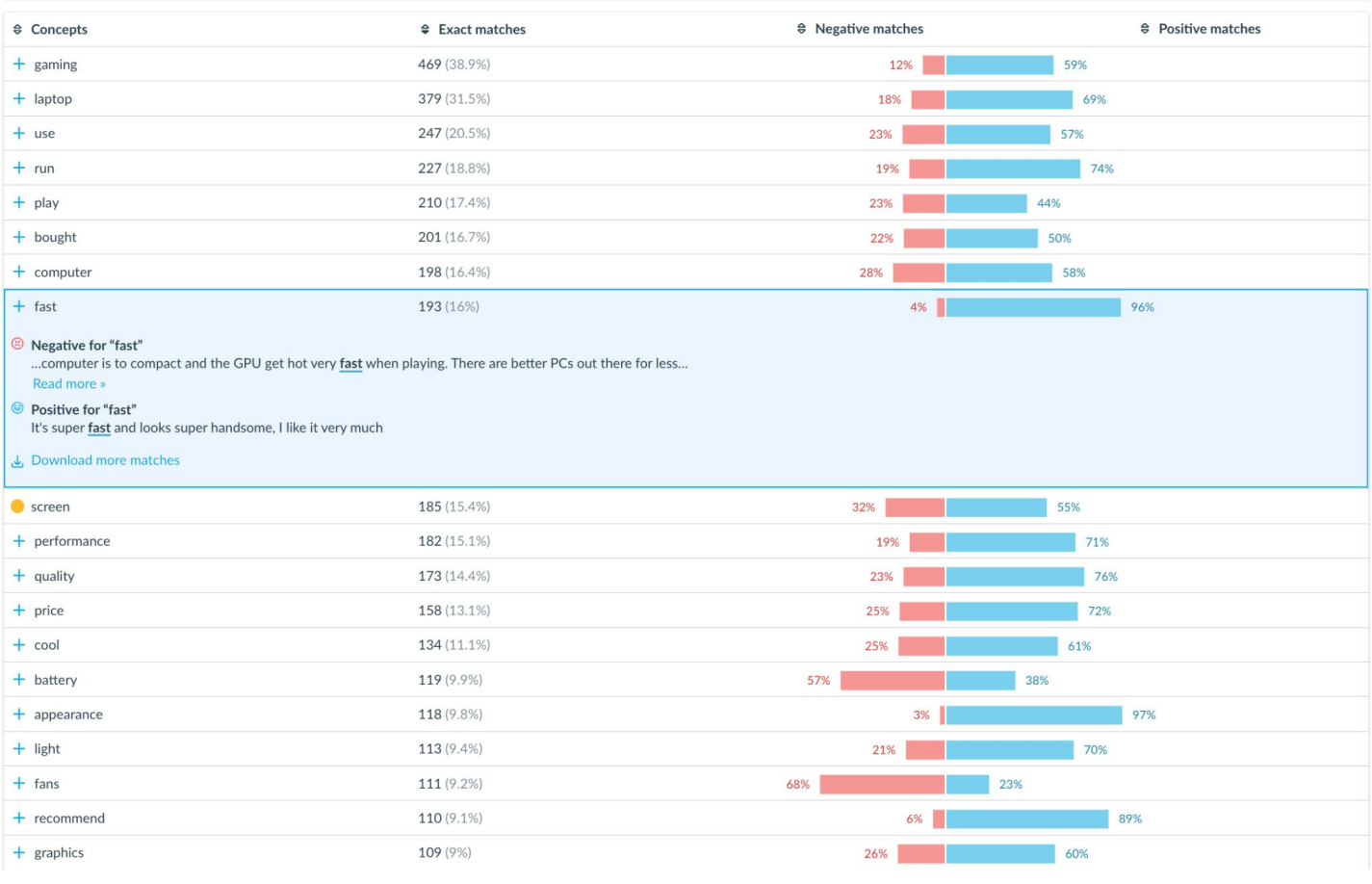


Over time measurements



Concepts	Total matches
taste	2,608 (38.4%)
melted_and_smell	2,552 (37.5%)
smell	2,256 (33.2%)
taste good	1,564 (23%)
daily vitamins	1,064 (15.7%)
delicious	1,050 (15.4%)
yummy	987 (14.5%)
bottle	728 (10.7%)
shipping	528 (7.8%)
packaging	492 (7.2%)
melted	387 (5.7%)

Concept level sentiment analysis



Concepts	Exact matches	Negative matches	Positive matches
+ gaming	469 (38.9%)	12%	59%
+ laptop	379 (31.5%)	18%	69%
+ use	247 (20.5%)	23%	57%
+ run	227 (18.8%)	19%	74%
+ play	210 (17.4%)	23%	44%
+ bought	201 (16.7%)	22%	50%
+ computer	198 (16.4%)	28%	58%
+ fast	193 (16%)	4%	96%
screen	185 (15.4%)	32%	55%
+ performance	182 (15.1%)	19%	71%
+ quality	173 (14.4%)	23%	76%
+ price	158 (13.1%)	25%	72%
+ cool	134 (11.1%)	25%	61%
+ battery	119 (9.9%)	57%	38%
+ appearance	118 (9.8%)	3%	97%
+ light	113 (9.4%)	21%	70%
+ fans	111 (9.2%)	68%	23%
+ recommend	110 (9.1%)	6%	89%
+ graphics	109 (9%)	26%	60%

Negative for "fast"
...computer is to compact and the GPU get hot very **fast** when playing. There are better PCs out there for less...
Read more »

Positive for "fast"
It's super **fast** and looks super handsome, I like it very much

Download more matches

Search enhancement using related concepts

Searchable text concept expansion from domain learning

Primary Text: Product Description text

...relief for skin conditions such eczema...

Domain conceptual matches to "**relief**"

Top related concepts

relief	1.00
relieve	0.69
reliever	0.66
soreness	0.62
alleviate	0.62
discomfort	0.61
pain	0.60

Domain conceptual matches to "**skin**"

Top related concepts

skin	1.00
epidermis	0.77
cutaneous	0.74
epidermal	0.73
Dermal	0.71
sun-damaged	0.68
not acne prone	0.67
invisibly	0.66
complexion	0.66

Domain conceptual matches to "**eczema**"

Top related concepts

eczema	1.00
dermatitis	0.84
psoriasis	0.82
scabies	0.75
atopic	0.74
excema	0.73
exzema	0.73
ezcema	0.72
hives	0.71
acne	0.70
seborrheic	0.69
pox	0.68
rash	0.68

Enhanced Indexing

Search is more flexible with additional matches possible

Primary Text: Exzana Product Description

Eczana is the all-natural solution to any skin condition resulting in irritated, scaly, itchy skin.

standard whitespace tokenizer

Searchable Text

Standard Tokanization Index

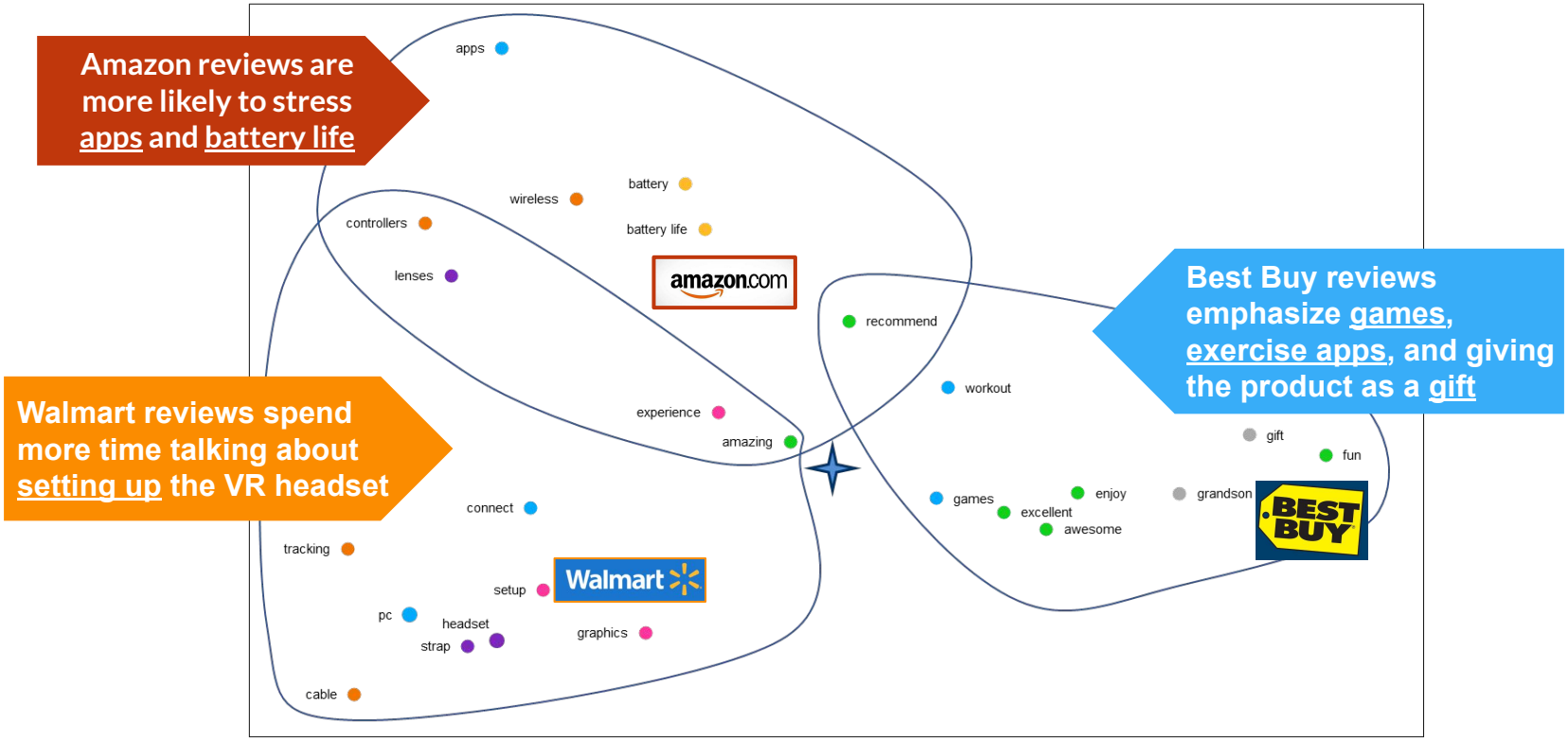
[Eczana, is, the, all-natural, solution, to, any, skin, condition, resulting, in, irritated, scaly, itchy, skin]

enhanced concepts

Index Enhancements

Eczana ->	Tazorac, topicals, rosacea,
All-natural ->	wholesome, safe-to-use, hand-crafted,
Solution ->	Panacea, solved, solved, remedy
Skin ->	Epidermis, cutaneous, epidermal, dermal
Condition ->	Conditional, situation
Resulting ->	Outcome, resultant, cumulative
Irritated ->	aggravate, irritation, inflamed, irritants
Scaly ->	itchy, flaky, peely, rashy, scaley

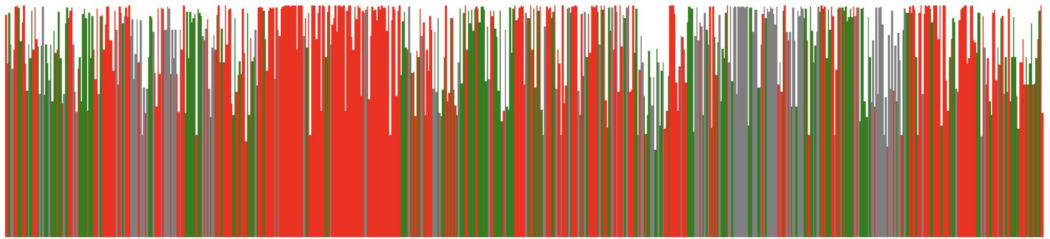
Leveraging metadata



https://www.mairesearch.com/
https://luminoso.com/products/luminoso-daylight/

Video transcript sentiment

anyone who tries to make a device to compete against the steam deck has a very tough time because this is a handheld gaming PC and it only costs 400 and a bunch of companies have tried right they've come out with devices that have had better Hardware than the steam deck but they've never had a compelling price but recently Asus came into the studio a nd they said we understand that this is a very competitive segment we understand that everyone loves that 400 price p oint of the steam deck but we have something that's even better the Rog Ally asus's new handheld gaming PC so I've sp ent a little over a week with this device and it is still just an early engineering sample but I think they got a sho t like I think this product does so much right it just makes for way better gaming experience than the steam deck le t's start off with performance so Asus claims that this has doubled the performance of the steam deck like literally 2X and I can't divulge into frame rate and stuff like I just took a little peek around they told me I can't share it but this is a brand new trip from AMD it's hot off their press it's a custom four nanometer Apu Zen 4 rdna3 now I do n't know the naming and the kind of clock rate on this thing but from what I've seen I think the claims are legit now arguably even more important than performance it well I guess it complements performance is the screen so this screen is a seven inch screen same size I guess as the screen on the steam deck but it's a way better screen so this is a hi gher res screen than the steam deck is 1080p instead of 800p it's significantly brighter has way better colors but th e greatest thing the display on the RG Ally goes up to 120 hertz it is such a different gaming experience when you're playing stuff with fluid visuals see last year when I reviewed the steam deck like this was impressive to me at the t ime to be able to get 50 or 60 frames per second but to be able to get that you had to be in 800p and for a lot of ga mes you have to be in low graphic settings but on the Rog Ally games will comfortably hit way higher frame rates whil e on higher graphic settings and for a lot of PC games it's a very different game experience with those higher frame rates I think this screen is awesome it's just that you need really good Hardware to be able to take advantage of it



Bias Detection Algorithm

Bias concepts will be outliers in the data

Start with the lowest prevalence concepts

Check those low prevalence concepts in each document where that low prevalence concept also has negative sentiment in context of the document.

Compare the concept to the list of common bias terms (man, woman, black, white, old guy, girl, gay, rainbow, etc)

POC Results

- 14 documents out of 5988 tagged as having bias
- The two control bias documents were found
- Some of the twelve false positives are also ‘interesting’

Interesting Bias False Positive 1

We found bias against ‘mold’ in a home inspection for a loan. The bias term in this email is “dark”. It’s still an interesting find and even though most humans ‘should’ have bias against black mold, we are finding bias wherever it exists.

> “Good afternoon Jim,
>
> Just wanted to touch base with you about Mold/Mildew Inspection in my house. I had an appointment with Stay Dry Waterproofing Ohio Company yesterday with an intention to diagnose the problem and to arrange an appropriate solution. Unfortunately, their Representative was not really interested in testing of the air or **dark** spots on my basement walls, he just wanted to sell a big cost products along with big item project. While still waiting for their estimate, I contacted another Company - Mold Test Company and made an appointment for 12/14/21 with them for testing the air and wall substance before to schedule remediation.“

Label Generation

Creating a classifier requires a labeled dataset

Create a list of concept lists that will be labels

Search for documents related to a specific concept

Dedupe documents in multiple labels

- Greetings and Wrap ups *(Adrian, Renee how can I help, bye bye, thank you bye by...*
- Inventory *(available, stock, availability)*
- Feature questions *(features, option, take a picture, recharge, charging station, road ...*
- Installation *(install, electrician to come, mount, screws, circuit breaker, junction box...*
- Offers *(offers, rebate, discount, tax credit, sale, earth day sale, sale price)*
- Partnership *(distributor, reseller, resale, wholesale)*
- Product Conversation *(prefer, juice box, juice box 32, juice box 40, G-Spot, juicebox)*
- Contact agent *(supervisor, department, customer service)*
- Shipping inquiry *(shipping, UPS, shipment, label created, delivery, tracking number,...*

label	string_Filena	match_score	number_Utt	text
Conversation: Wrap ups	608c4ed7a8	3.69574881	96	okay sounds good thank you bye bye
Conversation: Wrap ups	6092f1fb64b	4.30450678	46	okay sounds good alright take care bye bye
Conversation: Greetings	60872887dc	2.53509641	34	okay alright well then I'm going to let him kn
Conversation: Greetings	608c44dc292	2.38386822	24	okay great I appreciate it thank you very muc
Installation	60872f88b13	2.07760859	2	Adrian how you doing good afternoon this is M
Conversation: Greetings	608707bcb54	1.86525929	36	okay thank you very much Adrian
Conversation: Greetings	6084255bbd	1.85534501	40	you have a good day bye bye
Conversation: Greetings	6081e325dd	1.8509798	94	okay alright I think that's it I made a plan oka
Conversation: Wrap ups	6089d2d47d	2.10000014	32	thank you bye bye
Conversation: Wrap ups	60835455a4	2.10000014	54	thank you bye bye
Conversation: Wrap ups	6089eda6d7	2.10000014	35	okay alrighty thank you bye bye
Conversation: Wrap ups	6089c1010ef	2.10000014	29	alright thank you bye bye
Conversation: Wrap ups	6081dd5e52	2.10000014	32	alright thank you bye bye
Conversation: Wrap ups	6084a77967	2.09404421	33	YouTube thank you bye bye
Conversation: Wrap ups	608c424edf7	2.07927513	33	see you sure thank you bye bye
Shipping inquiry	60872887dc	3.35799861	2	hi Adrian this is Kerry from Costco.com just tc

LLMs and Vectorization

LLMs

LLMs offer amazing new capabilities

"*Well, the growth of AI in terms of weapons systems and the problems that it is going to create have been very apparent for a lot of years. Few journalists bothered to write about it. Now that there's a chatbot that can write an article for a local newspaper, suddenly it's a crisis.*"
Christopher Nolan - Wired - June 2023

https://www.wired.com/story/christopher-nolan-oppenheimer-ai-apocalypse/

LLMs and Context Windows

LLM capabilities are complementary to vectorization systems.

Why not just feed all my text into the LLM and have it do the processing?

Context windows are 4k to 100k. Not a lot of text

Let’s talk about context windows…

medium.com/@neonforge/chatgpt-api-and-python-how-to-simulate-linux-terminal-with-just-a-few-lines-of-code-35ab4dfbcab0

ChatGPT API and Python: How to Simulate Linux Terminal with just a Few Lines of Code

Michael King · Follow
4 min read · Mar 3

7

Can you believe what we're about to do next, my friend? We're going to show you how to simulate a Linux terminal using Python and OpenAI's ChatGPT API with just a few lines of code! It's like magic, only better, because it's real and totally achievable.

https://medium.com/@neonforge/chatgpt-api-and-python-how-to-simulate-linux-terminal-with-just-a-few-lines-of-code-35ab4dfbcab0

```python
import openai

openai.api_key = "YOUR_OPENAI_API_KEY"

impersonated_role = "I want you to act as a Linux terminal, " \
                    "I will type commands and you will reply with what the terminal should show. " \
                    "I want you to reply with the terminal output inside a unique code block and nothing else." \
                    "do not write explanations." \
                    "do not type commands unless I instruct you to do so." \
                    "When I need to tell you something in English I will do so by putting text inside curly brackets {something like this}.

command_history = ""
while True:
    command_prompt = input("Input command: ")
    output = openai.ChatCompletion.create(
        model="gpt-3.5-turbo",
        messages=[
            {"role": "system", "content": f"{impersonated_role}. based on this command history: {command_history}"},
            {"role": "user", "content": f"{command_prompt}"},
        ]
    )

    for item in output['choices']:
        print(item['message']['content'])
        command_history = f"{command_history}command:{command_prompt}\noutput:{item['message']['content']}\n"
```

https://medium.com/@neonforge/chatgpt-api-and-python-how-to-simulate-linux-terminal-with-just-a-few-lines-of-code-35ab4dfbcab0

````
Input command: pwd
```bash
/home/user
```
Input command: ls
output:```bash
Desktop  Documents  Downloads  Music  Pictures  Public  Templates  Videos
```
Input command: cd Documents
```bash
(no output, command changes current working directory to Documents)
```
Input command: ls
output:
```bash
file1.txt  file2.pdf  file3.docx  folder1  folder2
```
Input command: touch newfile.txt
output:```bash
(no output, command creates a new empty file named newfile.txt in the current directory)
```
Input command: ls
output:```bash
file1.txt  file2.pdf  file3.docx  folder1  folder2  newfile.txt
```
Input command:
````

https://medium.com/@neonforge/chatgpt-api-and-python-how-to-simulate-linux-terminal-with-just-a-few-lines-of-code-35ab4dfbcab0
````

````
Input command: top
```
top - 09:29:44 up 3 days,  7:54,  1 user,  load average: 0.00, 0.00, 0.00
Tasks: 110 total,   1 running, 109 sleeping,   0 stopped,   0 zombie
%Cpu(s):  0.0 us,  0.1 sy,  0.0 ni, 99.9 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
MiB Mem :  15802.8 total,    277.2 free,  11861.5 used,   3683.9 buff/cache
MiB Swap:      0.0 total,      0.0 free,      0.0 used.   2886.0 avail Mem

  PID USER      PR  NI    VIRT    RES    SHR S  %CPU  %MEM     TIME+ COMMAND
    1 root      20   0  169016   6416   4416 S   0.0   0.0   0:08.27 systemd
    2 root      20   0       0      0      0 S   0.0   0.0   0:00.02 kthreadd
    3 root      20   0       0      0      0 S   0.0   0.0   0:01.80 ksoftirqd/0
    4 root      20   0       0      0      0 S   0.0   0.0   0:00.01 kworker/0:0
```

Note: This is the output of the `top` command. The output is dynamic and may change every second.
Input command: ifconfig
```
eth0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 192.168.0.10  netmask 255.255.255.0  broadcast 192.168.0.255
        inet6 fe80::a00:27ff:fef9:3f3f  prefixlen 64  scopeid 0x20<link>
        ether 08:00:27:f9:3f:3f  txqueuelen 1000  (Ethernet)
        RX packets 81550  bytes 117447190 (111.8 MiB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 3480  bytes 792455 (773.6 KiB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

lo: flags=73<UP,LOOPBACK,RUNNING>  mtu 65536
        inet 127.0.0.1  netmask 255.0.0.0
        inet6 ::1  prefixlen 128  scopeid 0x10<host>
        loop  txqueuelen 1000  (Local Loopback)
        RX packets 151  bytes 13738 (13.4 KiB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 151  bytes 13738 (13.4 KiB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0
```
Input command:
````

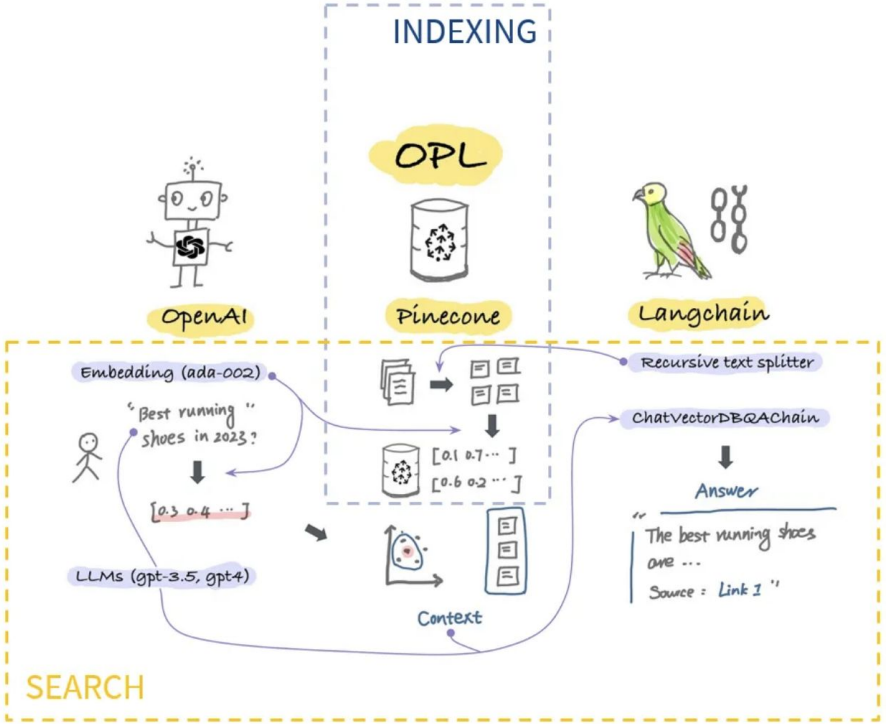
https://medium.com/@neonforge/chatgpt-api-and-python-how-to-simulate-linux-terminal-with-just-a-few-lines-of-code-35ab4dfbcab0

The context window issue

```
model="gpt-3.5-turbo",
messages=[
    {"role": "system", "content": f"{impersonated_role}. based on this command history: {command_history}"},
    {"role": "user", "content": f"{command_prompt}"},
]

for item in output['choices']:
    print(item['message']['content'])
    command_history = f"{command_history}command:{command_prompt}\noutput:{item['message']['content']}\n"
```

RAG - Retrieval Augmented Generation
VectorDb and LangChain solution



https://pub.aimind.so/a-holy-grail-of-text-ai-chatgpt-llm-generative-query-on-your-unlimited-custom-data-b31129082b9e

Past and Present

Symbolic AI - GOFAI

In the 80s and 90s we were going to solve everything with logic.

Question:
"Every fox is faster than some snail."

Solution:

$$\underbrace{\textit{For any fox } x}_{\forall x \in F}, \underbrace{\textit{there is a snail } y}_{\exists y \in S} \textit{ such that } \underbrace{x \textit{ is faster than } y}_{P(x,y)}$$

$$\forall x \in F, \exists y \in S, P(x, y)$$

Calcworkshop.com

https://calcworkshop.com/logic/predicate-logic/

UNDERSTANDING

We still don't have something that understands and learns.

The top companies are all using ML/Neural networks.

We still have deeply symbolic AI systems

Cyc (https://cyc.com/) and OpenCyc (https://slor.sourceforge.net/e_ocyc.htm)
Dbpedia (https://www.dbpedia.org/resources/ontology/)
ConceptNet (https://conceptnet.io/)

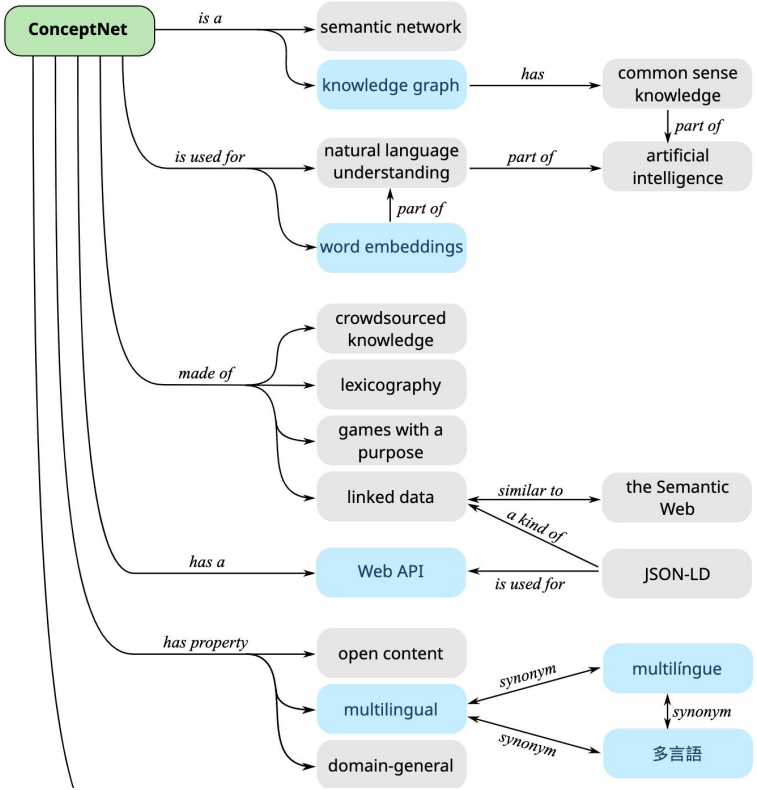
We've completely flipped from 1989 to 2023 where in 1989 there were very few people doing neural networks and everyone was doing GOFAI.

(opinion) We will need to combine the two for computers to really "understand" and manipulate data.

What is ConceptNet?

ConceptNet is a freely-available semantic network, designed to help computers understand the meanings of words that people use.

ConceptNet originated from the crowdsourcing project Open Mind Common Sense, which was launched in 1999 at the MIT Media Lab. It has since grown to include knowledge from other crowdsourced resources, expert-created resources, and games with a purpose.



Summary

Text Processing

The Vectorization Process

Applications of Vectorization

Neural Networks

LLMs

Symbolic AI

Thank you