91258 - Natural Language Processing

Lesson 3. Vector Space Model

Alberto Barrón-Cedeño

Alma Mater Studiorum-Università di Bologna
a.barron@unibo.it @_albarron_

09/10/2023

Table of Contents

Current Status

Representations Revisited

More Basic Algebra

Current Status

Current Status

You know. . .

- what is natural language processing
- there are two main paradigms: rule-based and statistical

On your own, you have. . .

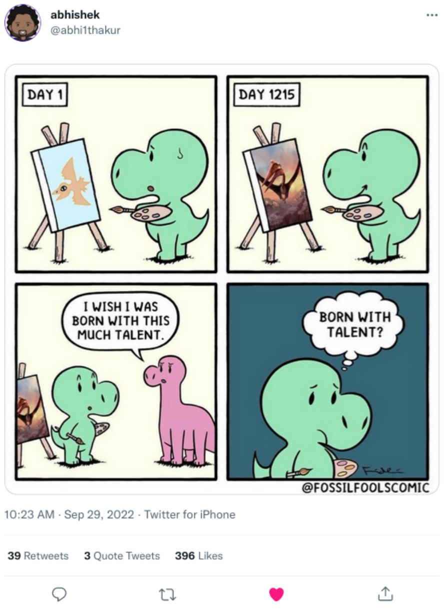
- setup a Python development environment
  1. command line
  2. PyCharm or any other option (e.g., Eclipse)
  3. Google's Colab

On your own, you (could) have. . .

- found out what is **git** (and perhaps LaTeX as well!)

You can. . .

- open a text file (Python intro)
- tokenise and normalise text
- build some text representations



abhishek
@abhi1thakur

10:23 AM · Sep 29, 2022 · Twitter for iPhone

39 Retweets 3 Quote Tweets 396 Likes

https://twitter.com/abhi1thakur/status/1575400771541155842

Representations Revisited

Representations Revisited

1. Use NLTK[1] or Spacy[2] to tokenize
2. Use `.lower()` to casefold (ignore capitalisation)
3. Use Porter's stemmer to drop suffixes
   or use a lemmatiser to find the *actual* root of words
4. Discard stopwords from the text*
5. Build a vectorial representation*

[1] https://www.nltk.org/

[2] https://spacy.io

Stopwords

Common words in a language that occur with a high frequency, but carry much less substantive information about the meaning of a phrase (Lane et al., 2019, p. 51–54)

Alternative 1 Consider the most frequent tokens in a reference corpus as stopwords (remember Genesis from P4P?)

Alternative 2 Take an existing list of stopwords[3]

en	es	it
i	a	altri
me	ahora	certa
my	alli	della
it	cerca	nessuna
is	el	prima
do	es	quello
the	unas	solito
will	vez	va
other	yo	via

[3] For instance, from NLTK, sklearn, or https://github.com/stopwords-iso

Stopwords

Discarding stopwords

- They are the most frequent tokens in the documents
- Discarding them reduces the computational effort significantly
- Typical size of a stopwords list: a few hundred words
- For some applications (e.g., **topic clustering**), they can be safely discarded
- For some others (e.g., **dialogue**) they cannot

Stopwords have to be considered with a grain of salt
(as everything in NLP)

Vector representation

BoW

- A text is represented as the bag (set) of its words
- It disregards grammar
- It disregards word order
- It (can) consider frequency

From (Lane et al., 2019, p. 41)

More Basic Algebra

x and y

Mathieu
@miniapeur

https://twitter.com/miniapeur/status/1710074831079690394

Dot product

Algebraically, it is the sum of the products of the corresponding entries of the two sequences of numbers $a \cdot b$

$$\begin{aligned} a \cdot b &= \sum_{i=1}^{n} a_i b_i \\ &= a_1 b_1 + a_2 b_2 + a_3 b_3 + \ldots + a_n b_n \end{aligned}$$

```
a = [1,2,3]
b = [3,4,6]
my_sum = 0
for i in range(len(a)):
    my_sum += a[i] * b[i]
```

There are better —more efficient— ways to compute the dot product!

Now, we can use the dot product to compare two documents ($\sim$ similarity)

Vector space model

"[. . .] an **algebraic** model for representing text documents (and any objects, in general) as vectors of identifiers [. . .]"[4]

Some applications

- Relevance rankings in keyword-based search
- Document clustering to "discover" structure and relations in a text collection

(not the SOTA for most tasks, but it's a *minimum viable product*)

</> **Let us see it working**

[4] https://en.wikipedia.org/wiki/Vector_space_model

Tomorrow. . .

VADER

References

Fernicola, F., S. Zhang, F. Garcea, P. Bonora, and A. Barrón-Cedeño
2020. Ariemozione: Identifying emotions in opera verses. In *Italian Conference on Computational Linguistics*.

Hutto, C. and E. Gilbert
2014. VADER:A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, Ann Arbor, MI.

Lane, H., C. Howard, and H. Hapkem
2019. *Natural Language Processing in Action*. Shelter Island, NY: Manning Publication Co.

Zhang, S., F. Fernicola, F. Garcea, P. Bonora, and A. Barrón-Cedeño
2022. AriEmozione 2.0: Identifying Emotions in Opera Verses and Arias. *IJCoL*, 8(2).