92586 Computational Linguistics

Lesson 4. Vector Space Model[1]

Alberto Barrón-Cedeño

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

05/03/2020

[1]Only notebooks were used in Lesson 3.

Table of Contents

Current Status

Current Status

You know...

- what is computational linguistics / natural language processing
- there are two main paradigms: rule-based and statistical
- how to identify the rule-based paradigm

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 (from P4P)
- tokenise and normalise text
- build some text representations

Wondering about your final project?

Get inspiration from last year's:

`albarron.github.io/teaching/computational-linguistics2020/`

Why not starting with a fantastic LaTeX template?

`github.com/TinfFoil/learning_dit_coli_projecttemplate`

Representations Revisited

Representations Revisited

1. Use NLTK[2] to tokenize
2. Use `.lower()` to neglect capitalisation
3. Use Porter's stemmer to drop suffixes
4. Or use a lemmatiser to find the root of the words
5. Build a vectorial representation
6. **Discard stopwords**

[2] `http://www.nltk.org/`

Stopwords

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

Many alternatives exist

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

[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
- Stopwording makes sense for some applications (e.g., **topic clustering**),
- It does not for some others (e.g., **dialogue**)

Stopwords have to be considered with a grain of salt (as most 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)

Dot product

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

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

```
v1 = [1,2,3]
v2 = [3,4,6]
my_sum = 0
for i in range(len(v1)):
    my_sum += v1[i] * v2[i]
```

(there are better ways to compute the dot product)
Now, we can compare documents ($\sim$ similarity)

Vector space model

We just built our first 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
- Text clustering to "discover" structure and relations in a text collection
- Reading recommendations

(Not the SoA for most tasks, but it's a starting point)

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

Sentiment Analysis

Sentiment Analysis

It **does not** refer to real sentiment, such as love or hate

It is about **positive** and **negative** (and **neutral**)

This monitor is definitely a good value. Does it have superb color and contrast? No. Does it boast the best refresh rate on the market? No. But if you're tight on money, this thing looks and preforms great for the money. It has a Matte screen which does a great job at eliminating glare. The chassis it's enclosed within is absolutely stunning.

POSITIVE

His [ssa] didnt concede until July 12, 2016. Because he was throwing a tantrum. I can't say this enough: [kcuF] Bernie Sanders.

NEGATIVE

From (Lane et al., 2019, p. 62–65)

Sentiment Analysis

VADER a rule-based approach

Valence **A**ware **D**ictionary for s**E**ntiment **R**easoning[5]

- It has a lexicon packed with tokens and their associated "sentiment" score
- It counts all tokens belonging to each category: [positive, neutral, negative]

[5] http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf

Coming next...

Statistical NLP

References

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