Lab 1 Questions and Answers

Guillaume CARRIERE, Paul GROLIER,
Cloé ESCUDIER

Contents

1 Keywords Extraction 2

1.1 Question 1.3 . . . 2

1.2 Question 1.4 . . . 2

1.3 Question 1.5 . . . 2

1.4 Question 2.2 . . . 2

2 Word Vectors 3

2.1 Question 2.5 . . . 3

3 Prediction-based word vectors 3

3.1 Question 3.1 . . . 3

3.2 Question 3.2 . . . 4

3.3 Question 3.3 . . . 4

3.4 Question 3.4 . . . 4

3.5 Question 3.5 . . . 4

3.6 Question 3.6 . . . 4

3.7 Question 3.7 . . . 4

3.8 Question 3.8 . . . 5

3.9 Question 3.9 . . . 5

4 Prediction-based sentence vectors 5

4.1 Question 4.1 . . . 5

4.2 Question 4.2 . . . 5

4.3 Question 4.3 . . . 5

4.4 Question 4.4 . . . 6

1 Keywords Extraction

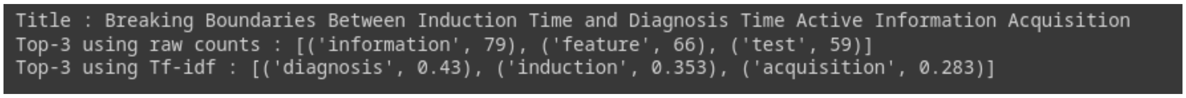
1.1 Question 1.3

One limit is that the most frequent word in a document might actually be common words which are very frequent in all documents, in the same manner as stop words. One could improve this approach by adding these words in our stop words list.

1.2 Question 1.4

We can test multiple values of max_df and pick the one with the best performance. Performance can be evaluated depending on our objective, for example it can be manual, or evaluated on a semantic search task. We are using sparse matrix because most of the rows will be equal to 0 as most of the words in our vocabulary will not appear in the text.

1.3 Question 1.5



```
Title : Breaking Boundaries Between Induction Time and Diagnosis Time Active Information Acquisition
Top-3 using raw counts : [('information', 79), ('feature', 66), ('test', 59)]
Top-3 using Tf-idf : [('diagnosis', 0.43), ('induction', 0.353), ('acquisition', 0.283)]
```

On this example, the top-3 words are "information, feature, test" using raw counts and "diagnosis, induction, acquisition" using tf-idf. In this case the top-3 words using raw counts seem to be more generic than the top-3 words using tf-idf (these appear in the title of the document). Thus, we would use the tf-idf as it seems more likely to output the correct document in a context of semantic search.

1.4 Question 2.2

We define our quality score for each metric as : $\mathsf{keywords\_quality} - 0.5 * \mathsf{computation\_time}$. We want to give more importance to keyword quality than computation time. We will also normalize these internal metrics.

We define $\mathsf{keywords\_quality}$ as follows :

$$\mathsf{keywords\_quality} = \sum_{d \in C} \sum_{w \in top10} \mathbb{1}_T(w, d) * (10 - \mathsf{rank}(w))$$

with C the corpus of documents, $\mathsf{top10}$ the top-10 keywords returned by the method and $\mathsf{rank}(w)$ the rank of a word in the top-10. Finally, with T_d the set of words appearing in the preprocessed title of a document d, we have:

$$\mathbb{1}_T(w, d) := \begin{cases} 1 & \text{if } \ w \in T_d \ , \\ 0 & \text{if } \ w \notin T_d . \end{cases}$$

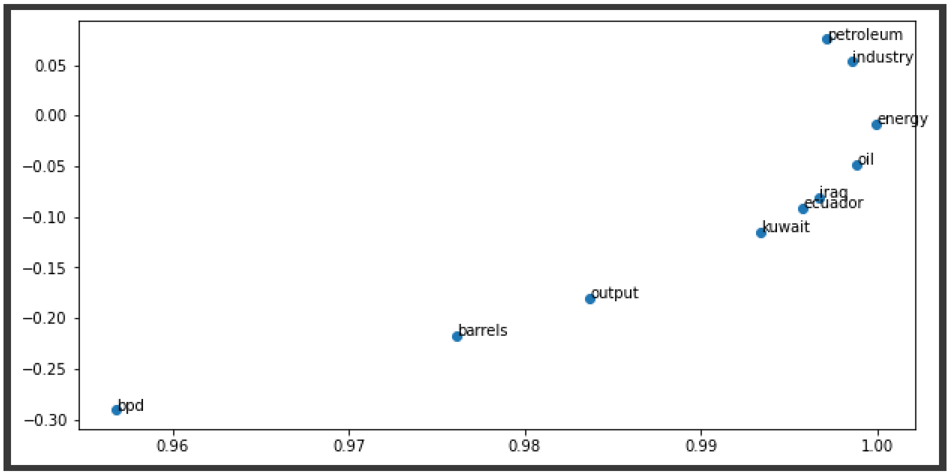
(note that $w \in T_d$ also if w appears as a subword of a word in T_d)

Méthode	Quality score	Time	Time optimisations
Raw counts	0.365	11 s	- Use a faster sorting algorithm for the counts
Td-idf	0.322	34 s	- Hyperparameter tuning on the count vectorizer
KeyBert	- 0.187	259 s	- Feeding multiple documents simultaneously (instead of one-by-one), as it speeds up the embedding - Improve the GPU's specs

According to our quality score the best method remains raw counts, very close to td-idf. Tough maybe our metric for deciding the quality of the extracted keywords is not good. We decided keywords were relevant if they appeared in the title, as we had some sort of document retrieval task in mind, but it is possible that keywords appearing in the title may not make for a good representation of the document as a whole. For example, these keywords could not help us to guess similarity of documents. In this kind of context, KeyBERT, which is based on BERT, would probably perform better.

2 Word Vectors

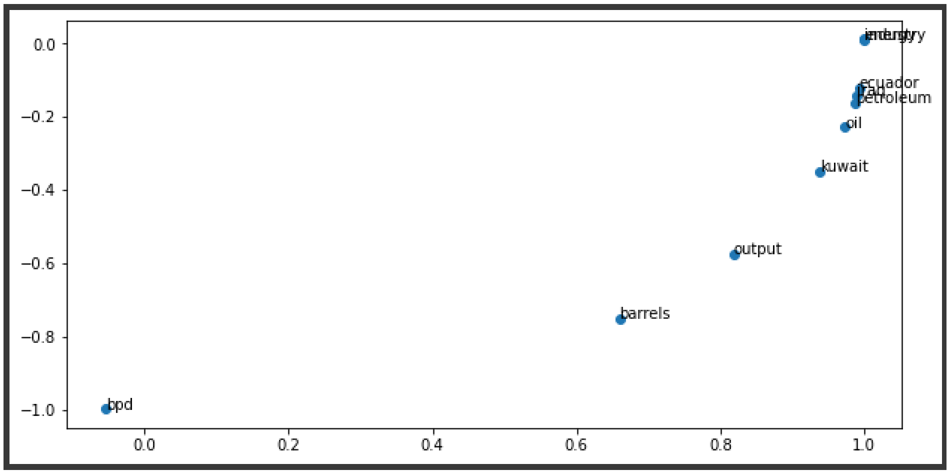
2.1 Question 2.5



We have a clear cluster for country names, such as "iraq, ecuador, kuwait" which is normal. We can also see that "petroleum" and "industry" are very close, which makes sense as "petroleum industry" is a common phrase. However we see that "bpd (barrels per day)" and "barrels" are far apart even though intuitively they should be close, both referring to barrels. This means they may actually be used in different contexts.

3 Prediction-based word vectors

3.1 Question 3.1



Now the main clusters are "iraq, ecuador, petroleum" and "industry, energy". "barrels" and "bpd" still do not form a cluster, confirming our previous intuition. This plot is different from the previous one, for example "industry" is now closer to "energy" than to "petroleum". This may be explained to the fact that GloVe is finer in learning his embedding to represent semantic meaning from the co-occurence matrix. The dataset used for the training of GloVe is probably way bigger and more generalist than the reuters corpus, which may also explain the differences (for example, "energy industry" may be more frequent than "petroleum industry").

3.2 Question 3.2

When using "mouse" we can see that in the top-10 words some relate to "mouse" as an animal ("mice", "rat", "rabbit", "mickey") and some relate to "mouse" as the device to use a computer ("keyboard", "cursor", "trackball", "clicks"). Many polysemous and homonymic words don't show this tendency on the first top-10 words. This may be due to the fact that one meaning is often way more frequent than the other on polysemous and homonymic words, which would lead for the more frequent meaning to have a bigger representation in the co-occurence matrix and a clearer presence in the words with best similarity.

3.3 Question 3.3

This counter-intuitive example may be explained by the fact that "confidential" is not as frequent as "private" and "public". These two words are also very frequently mentioned together as they are typical opposites (such as up and down, high and low, hot and cold).

3.4 Question 3.4

The expression with which we want to maximise the cosine similarity of x is $w + k - m$, so we want to maximise the similarity with woman and king and minimise it with man.

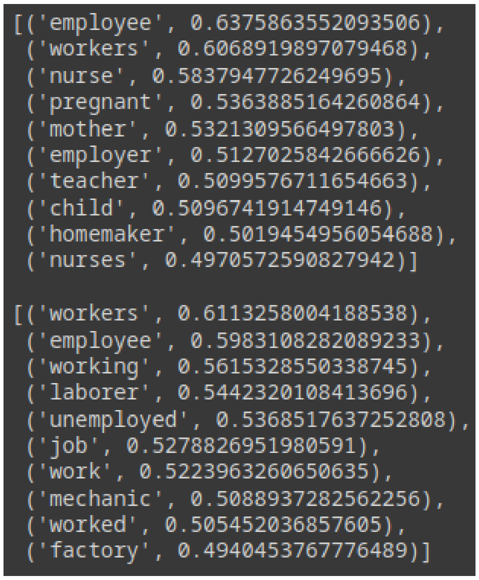
3.5 Question 3.5

The analogy is "eating : food :: drinking : water".

3.6 Question 3.6

The analogy is "bird : sky :: fish : rainbow".

3.7 Question 3.7

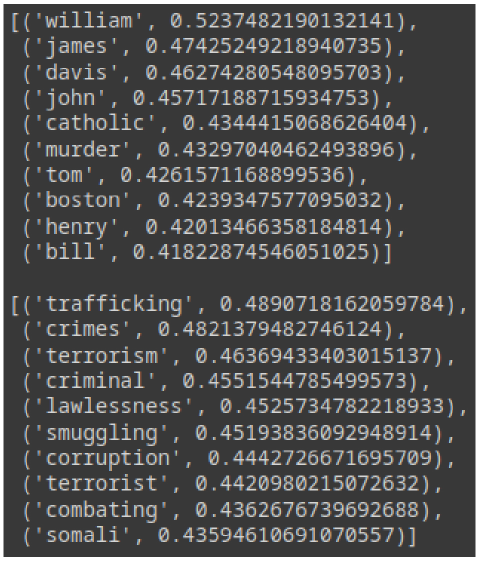


```
[('employee', 0.6375863552093506),
 ('workers', 0.6068919897079468),
 ('nurse', 0.5837947726249695),
 ('pregnant', 0.5363885164260864),
 ('mother', 0.5321309566497803),
 ('employer', 0.5127025842666626),
 ('teacher', 0.5099576711654663),
 ('child', 0.5096741914749146),
 ('homemaker', 0.5019454956054688),
 ('nurses', 0.4970572590827942)]

[('workers', 0.6113258004188538),
 ('employee', 0.5983108282089233),
 ('working', 0.5615328550338745),
 ('laborer', 0.5442320108413696),
 ('unemployed', 0.5368517637252808),
 ('job', 0.5278826951980591),
 ('work', 0.5223963260650635),
 ('mechanic', 0.5088937282562256),
 ('worked', 0.505452036857605),
 ('factory', 0.4940453767776489)]
```

We can see when "man" is positive and "woman" is negative we have words related to physical work such as "laborer, mechanic, factory", whereas when "woman" is positive and "man" is negative we have words related to medical or social work or housewives such as "nurse, teacher, cihld, homemaker". It reflects the stereotypes about men doing physical work and women staying at home or doing only "caring" work.

3.8 Question 3.8



```
[('william', 0.5237482190132141),
 ('james', 0.47425249218940735),
 ('davis', 0.46274280548095703),
 ('john', 0.45717188715934753),
 ('catholic', 0.4344415068626404),
 ('murder', 0.43297040462493896),
 ('tom', 0.4261571168899536),
 ('boston', 0.4239347577095032),
 ('henry', 0.42013466358184814),
 ('bill', 0.41822874546051025)]

[('trafficking', 0.4890718162059784),
 ('crimes', 0.4821379482746124),
 ('terrorism', 0.46369433403015137),
 ('criminal', 0.4551544785499573),
 ('lawlessness', 0.4525734782218933),
 ('smuggling', 0.45193836092948914),
 ('corruption', 0.4442726671695709),
 ('terrorist', 0.4420980215072632),
 ('combating', 0.4362676739692688),
 ('somali', 0.43594610691070557)]
```

The first list of words are related positively to "thomas, crime" and negatively to "mohamed", and on the second list of words "thomas" is negative and "mohamed, crime" is positive. We can see the first list has mostly generic other occidental names ("william", "james", "john"), while the second list contains words related to terrorism and crime, such as ("trafficking", "terrorism", "criminal"). This show a bias reflecting the stereotype of arabic names being more related to crime than occidental names.

3.9 Question 3.9

This bias can come from the fact the dataset contains these stereotypes. Thus it is easier for the model to exploit the stereotypes to maximise its similarities when learning the embedding. To test and measure this we could try to modify the corpus and see how these biases evolve to see how strong they are.

4 Prediction-based sentence vectors

4.1 Question 4.1

We could represent a sentence by computing the average of the embeddings of each word in the sentence. However this would give an equal weight to each word in the sentence even though some words carry more meaning in the sentence. This would also fail to relate link between specific words or exploit particular n-grams with high meaning.

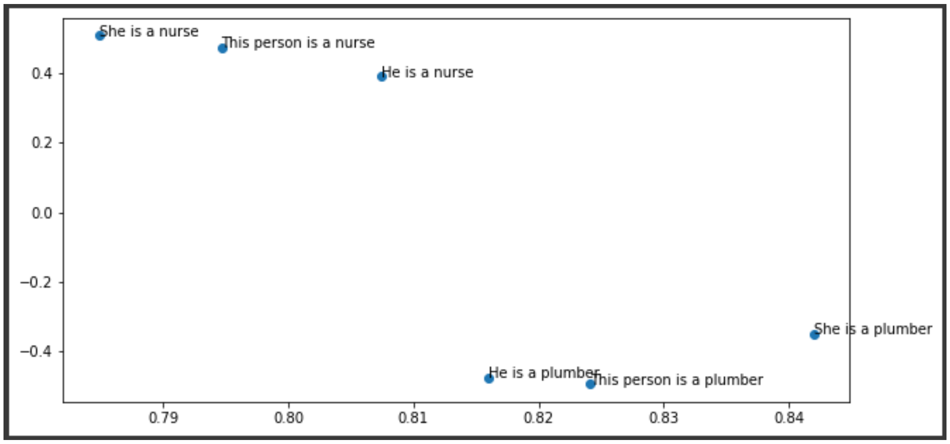
4.2 Question 4.2

The clustering quality on this sample is very good as the most similar sentences have been detected. This is because the sentences are embedded as a whole leading to efficient semantic text similarity using cosine similarity. The method we would choose is agglomerative clustering, as it seems to be better when the number of clusters is not known in advance.

4.3 Question 4.3

We can see that sentences with the same meaning are almost on the same coordinates. Sentences with very similar meanings still form tight clusters (like the cluster formed by the sentences in the form "A man is eating ***"). We can also see that clusters with similar words or words which are seen frequently together are close, for example the clusters with sentences containing "monkey, gorilla", and "cheetah" are close. The clusters which begin with "A man" are also close in the same fashion.

4.4 Question 4.4



In this examples, we replace "This person" by "she" or "he" to see if the model tends to relate an activity to the male or female gender. By doing this, we can see the model associates the activity of "nurse" to the female gender more than the male gender, and associates the activity of "plumber" to the male gender more than the female gender". This show a bias that reflects the gender stereotype.