Barcelona GSE Data Science Center

Text Mining for Social Sciences

Nandan Rao

April, 2019

Outline

- Informatiion Retrieval
- NLP
- Preprocessing Preprocessing Preprocessing

Social Sciences meets ML?

When solving a problem of interest, do not solve a more general problem as an intermediate step. Try to get the answer that you really need, but not a more general one. (Vladimir Vapnik)

Statistical Modelling: The Two Cultures

https://projecteuclid.org/download/pdf_1/euclid.ss/1009213726

What is Information Retrieval?

- Information retrieval ≈ search.
- One of the basic, early problems of internet engineering and information organization.
- Many of the tools we use in NLP were created for this problem.
- You have a corpus of documents (for example: the internet). You have a user who wants a few of these documents. How do you design this system?

Information Retrieval - Naive Search

Let's say you are inventing search. Imagine someone searching for the term "People who see ghosts". How could you pick between the following?

- This is a document about people who see ghosts. Those people end up on TV shows.
- This is a document about seeing goats. Those people work on farms.

Information Retrieval - Naive Search

Let's try again with the term: "People who see ghosts"

"I don't believe people who see ghosts", said Mannie, before spitting into the wind and riding his bike down the street at top speed. He then went home and ate peanut-butter and jelly sandwiches all day. Mannie really liked peanut-butter and jelly sandwiches. He ate them so much that his poor mother had to purchase a new jar of peanut butter every afternoon.

We have collected a report of every resident in our community that has seen a ghost. Each resident was asked "how many ghosts have you seen?", "describe the last ghost you saw", and "tell us about your mother." Afterwards, we compared the ghost reports between the different individuals, and assessed whether or not they had actually seen these apparitions.

Information Retrieval - Naive Search

Let's try again with the term: "People who see ghosts"

"I don't believe **people who see ghosts**", said Mannie, before spitting into the wind and riding his bike down the street at top speed. He then went home and ate peanut-butter and jelly sandwiches all day. Mannie really liked peanut-butter and jelly sandwiches. He ate them so much that his poor mother had to purchase a new jar of peanut butter every afternoon.

We have collected a report of every **resident** in our community that has **seen** a **ghost**. Each **resident** was asked "how many **ghosts** have you **seen**?", "describe the last **ghost** you **saw**", and "tell us about your mother." Afterwards, we compared the ghost reports between the different **individuals**, and assessed whether or not they had actually **seen** these **apparitions**.

Information Retrieval - Term Frequency

- Frequency matters!
- Let’s try and count the frequency of each word

Information Retrieval - Linguistic Tricks

Stop words “seen a ghost” → “seen ghost”

Stemming “seen a ghost” → “see ghost”

Lemmatization “saw ghosts” → “see ghost”

Tokenization “see ghost” → [“see”, “ghost”]

Information Retrieval - Synonyms

We might need some concept of synonyms.

- ghost, apparitions, spook $\rightarrow$ ghost
- people, individuals, residents, folk $\rightarrow$ people

Are these actually synonyms?

Information Retrieval - TF Fail

Now let's try our tools on the following text:

> *People see incredible things. One time I saw some people talking about things they had seen, and those people were so much fun. They saw clouds and they saw airplanes. Can you believe the amount of seeing done by these people? People are the best.*

Information Retrieval - IDF

Let df_v be the number of documents that contain the term v.

The *inverse document frequency* is

$$\text{idf}_v = \log\left(\frac{D}{df_v}\right),$$

where D is the number of documents.

Properties:

1. Higher weight for words in fewer documents.
2. Log dampens effect of weighting.

Information Retrieval - IDF

For words which are more common, we lower their weights.

(example)

Information Retrieval - IDF

Words which appear in *many* of the documents are not going to help us pick *one* document.

Natural Language Processing

What is Natural Language Processing?

- https://en.wikipedia.org/wiki/Natural-language_processing#History
- https://www.cl.cam.ac.uk/archive/ksj21/histdw4.pdf

Natural Language Processing

Two large challenges of Natural Language Processing:

- Put language into a metric space.
- Deal with the complex correlations between words in a sentence, and sentences in a document.

Metric Space

A metric space consists of a set (we'll call them **documents** in this context) and a distance metric between items in the set.

- What are some possible measures of "distance" between two documents?

Space

(word count example)

Distances

Now that we have our data into a numeric form, how can we determine a distance?

Distances

What about the distance between these two documents?

- We have collected a report of every resident in our community that has seen a ghost. Each resident was asked “how many ghosts have you seen?”, “describe the last ghost you saw”, and “tell us about your mother.” Afterwards, we compared the ghost reports between the different individuals, and assessed whether or not they had actually seen these apparitions.
- We ask each resident how many ghosts they’ve seen.

Distances

We *might* want a distance that ignores the "size" of the document.

One option is to normalize our vectors to unit length, this has the advantage of keeping the "direction" while removing the size element. Once we normalize our vectors, the euclidian distance becomes proportional to:

$$||A - B||^2 \propto 1 - A^T B = 1 - \cos\theta$$

Where $1 - \cos\theta$ is referred to as the **cosine distance** and $1 - \cos\theta$ is referred to as the **cosine similarity**.

Space

What about the similarity of these two documents:

People who see ghosts are full of crap. I don't believe a word they say. They didn't actually see any ghosts. No way! They are just seeing things.

We talked to lots of people who have seen ghosts. Each person was asked "how many ghosts have you seen?" They had a lot of interesting and disturbing stories about the ghosts in their lives.

NGrams

With the previous example, stemming/lemmatization + NGrams + TF-IDF would yield a feature:

“see ghost”

which would most likely be very highly weighted (depending on the corpus). This would help these two documents to be very similar, even though they are not in the simple BOW space.

NGrams

Similarly, in a 2-gram space these two documents separate:

Sometimes, at my job, I use text mining.

Sometimes, at my mining job, I text.

NGrams

"at my job, I use text mining"

```
["at my", "my job", "job I",
"I use", "use text", "text mining"]
```

"at my mining job, I text"

```
["at my", "my mining", "mining job",
"I text"]
```

Distances

What about a continuous metric space?

Distances

Co-occurences can be used as a proxy for semantic similarity.

Distances

(embedding example)

Distances

Other ways to think about semantic structure of a sentence?

Grammar???

Natural Language Processing

- In an attempt to create conversations, computer scientists brought in linguists.
- There was a need to understand the semantic content of sentences.

Natural Language Processing

How can we differentiate between these documents?

- France: Migrant stabbed to death in Calais
- Afghan asylum seeker stabbed to death in London park
- Clashes in Istanbul after angry mourners of a Turkish man is stabbed to death by an Afghani refugee
- German woman stabbed to death by Syrian refugee on her doorstep
- In memory to Bangladeshi migrant #Manan stabbed to death 6y ago during pogrom orchestrated by Nona's
- great people? the people that kicked over jugs of water to let migrants die in the desert? those are not great people.

Machine Learning with Language

https://github.com/nandanrao/text-mining/blob/master/dependency-tree-example.ipynb

Barcelona GSE
Data Science Center