Text Mining for Social Sciences

Nandan Rao

April, 2019

- Informatiion Retrieval
- NLP
- Preprocessing Preprocessing Preprocessing (Workshop)

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 Modeling: The Two Cultures

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

- 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?

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.

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.

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.

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

Stop words "seen a ghost" → "seen ghost"

Stemming "seen a ghost" → "see ghost"

Lemmatization "saw ghosts" → "see ghost"

Tokenization "see ghost" → ["see", "ghost"]

We might need some concept of synonyms.

- ghost, apparitions, spook → ghost
- people, individuals, residents, folk → people

Are these actually synonyms?

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.

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

The *inverse document frequency* is

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

where D is the number of documents.

Properties:

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

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

(example)

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

What is Natural Language Processing?

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

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.

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

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

\url{https://github.com/nandanrao/text-mining/blob/master/Dependency-Tree%20Example.ipynb}

- Introduce Libraries
- Ngrams
- Stemmers & Lemmatizers
- Wordnet & Synonyms
- Sparse Matrices
- Language Detection / Multilingual
- Vector Embeddings