# Text Mining for Social Sciences (Introduction)

**Nandan Rao**
nandan.rao@upf.edu

# Outline

- Motivation and examples in the social sciences
- Informatiion Retrieval
- NLP
- Kaggle Competition
- Preprocessing Preprocessing Preprocessing (Workshop)

# Social Sciences?

What is text mining and why is it useful in the Social Sciences?

# Social Sciences - Predicting Political Violence

"Reading Between the Lines: Prediction of Political Violence Using Newspaper Text" - Hannes Mueller, Christopher Rauh

- Use newspaper text to predict when conflict will arise in a given country.
- Prediction is important for policy interventions.
- "We find, for example, that news that describes judi- cial procedures systematically decreases before conflict occurs."

# Social Sciences - Effects of Central Bank Communication

"Shocking language: Understanding the macroeconomic effects of central bank communication" - Stephen Hansena, Michael McMahon

- Central banks set policy, but they also communicate (forward guidance)
- LDA to generate 15 topics, pick 5 relevant topics, use in FAVAR regression.
- Create list of words deemed expansionary/contractionary. Count how many of these words show up in each topic. Use these as features in VAR.

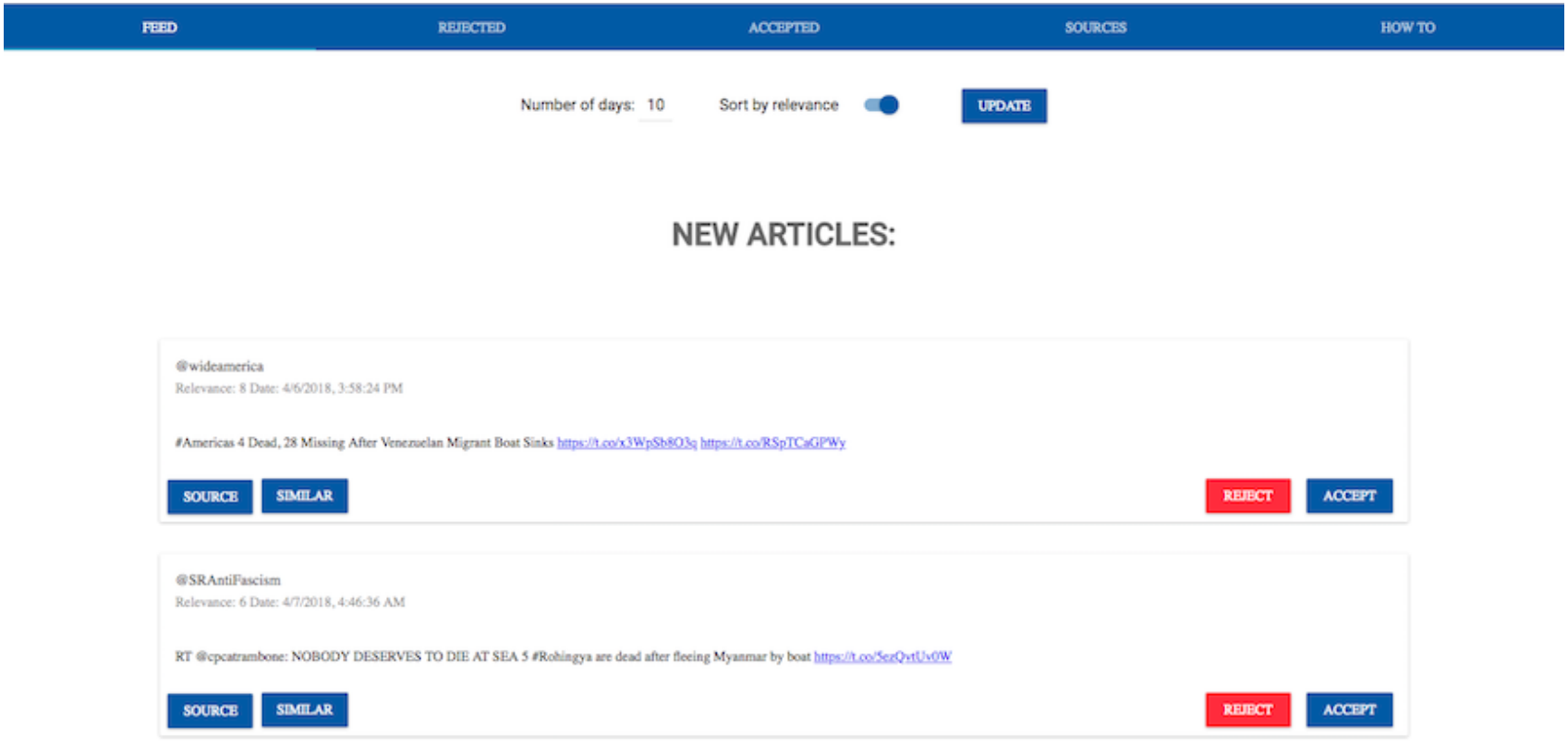
# Social Sciences - Missing Migrants


Missing Migrants Project tracks incidents involving migrants, including refugees and asylum-seekers, who have died or gone missing in the process of migration towards an international destination.

# Social Sciences - Missing Migrants

- The Missing Migrants project is essentially one of dataset creation.
- These statistics are not available from any single state.
- News agencies report the events. In plain text!

# Social Sciences - Missing Migrants



# Social Sciences - OECD & Freelance Labor Markets

What is the effect of the explosion of freelancing websites on the labor market? Demand side:

- "I need an experienced Business Strategist who can write content explaining all the important moving parts and pieces of building a business plan and/or business model. You'll be explaining to first time entrepreneurs and small business owners and diving into the importance"
- "We are Ricardo Steak House Restaurant located in Harlem, New York. We are looking for an expert opinion and training on how to manage our accounting department"
- "We are an 8Mil per year trucking company based out of NJ. Due to negative loss-runs, we lost ideal market coverage for insurance and forced to use Progressive Commercial. We need someone with both an accounting background and deep knowledge of commercial insurance..."

# 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)*

https://projecteuclid.org/download/pdf_1/euclid.ss/1009213

# What is Information Retrieval?

- Information retrieval $\approx$ 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 people who see 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” $\rightarrow$ “seen ghost”

# Information Retrieval - Linguistic Tricks

**Stop words** “seen a ghost” $\rightarrow$ “seen ghost”
**Stemming** “seen a ghost” $\rightarrow$ “see ghost”

# Information Retrieval - Linguistic Tricks

**Stop words** “seen a ghost” $\rightarrow$ “seen ghost”
**Stemming** “seen a ghost” $\rightarrow$ “see ghost”
**Lemmatization** “saw ghosts” $\rightarrow$ “see ghost”

# Information Retrieval - Linguistic Tricks

**Stop words** “seen a ghost” $\rightarrow$ “seen ghost”
**Stemming** “seen a ghost” $\rightarrow$ “see ghost”
**Lemmatization** “saw ghosts” $\rightarrow$ “see ghost”
**Tokenization** “see ghost” $\rightarrow$ [“see”, “ghost”]

# Information Retrieval - Synonyms

We might need some concept of synonyms.

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

# 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

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

where $D$ is the number of documents.

Properties:

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

# 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.

# 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

# Machine Learning with Language

https://github.com/nandanrao/text-mining/blob/master/Depend
20Example.ipynb

# Workshop

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