AMPHAN

NLP Overview

DSSG SOLVE

IWMI
International Water
Management Institute

Solve Team

Ancil Crayton

Project Manager

Jared Ross

Data Scientist

João Fonseca

Data Scientist

Kanav Mehra

Data Scientist

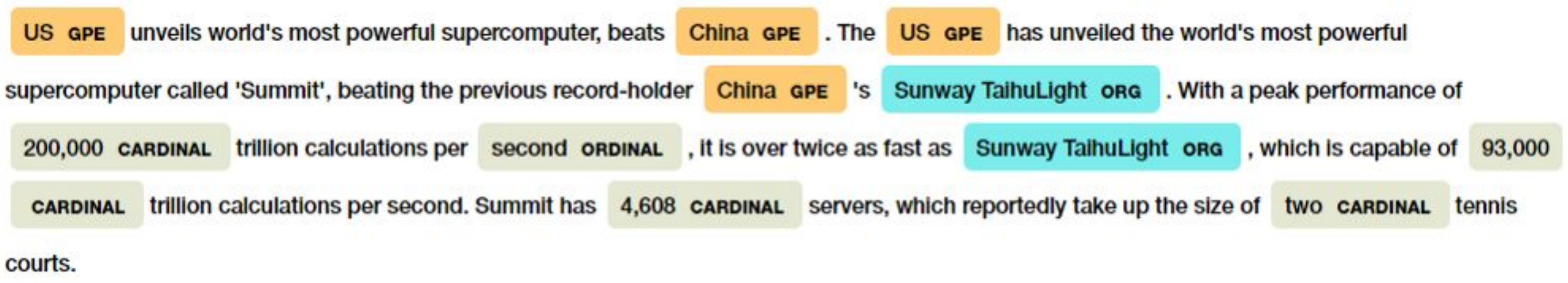
Marcelo Sandoval-Castañeda

Data Scientist

Introduction

“Natural language processing is a field of artificial intelligence that gives [the] machines the ability to read, understand and derive meaning from human languages.”

Named Entity Recognition (NER)



US GPE unveils world's most powerful supercomputer, beats China GPE . The US GPE has unveiled the world's most powerful supercomputer called 'Summit', beating the previous record-holder China GPE 's Sunway TaihuLight ORG . With a peak performance of 200,000 CARDINAL trillion calculations per second ORDINAL , it is over twice as fast as Sunway TaihuLight ORG , which is capable of 93,000 CARDINAL trillion calculations per second. Summit has 4,608 CARDINAL servers, which reportedly take up the size of two CARDINAL tennis courts.

“Named-entity recognition (NER) (also known as entity identification, entity chunking and entity extraction) is a subtask of **information extraction** that seeks to locate and **classify named entities** mentioned in unstructured text into **predefined categories** such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc.”

Popular methods:

Lexicon approach

Rule-based systems

Machine learning-based systems

Hybrid approach (ML + rules)

Application of NER

Motivation: Extract entities from tweets to better track the roles of individuals and organizations

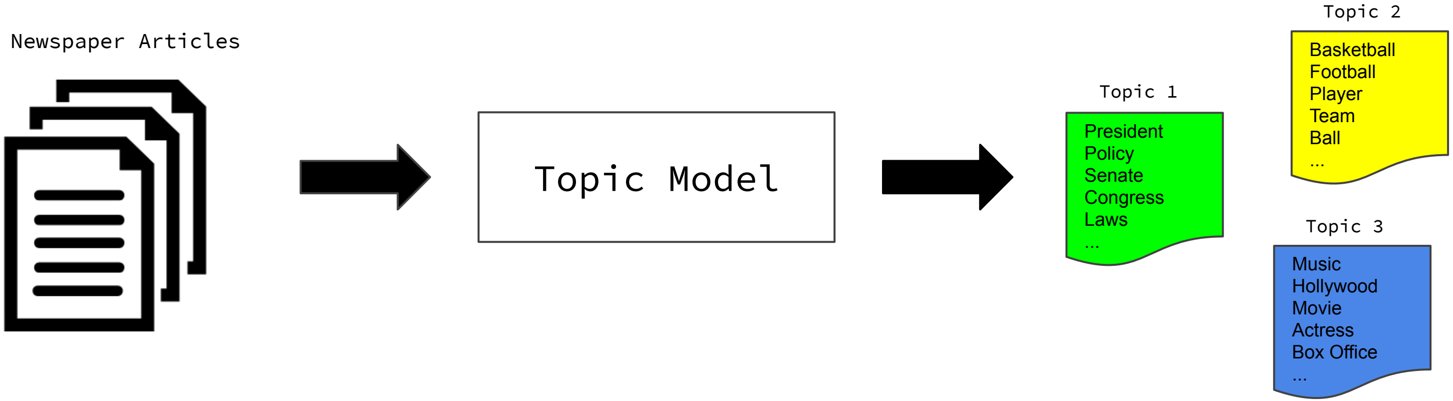
Example 1: Extract mentions of organizations or individuals and rank mentions to get a sense of who is influential

Example 2: Combine NER with entity linking to construct a knowledge graph that allows a user to query relevant information present in natural disaster discourse

Part of Speech Tagging

Application of Part of Speech Tagging

Topic Modeling



“A topic model is a type of **statistical model** for discovering ‘abstract’ topics that occur in a collection of documents”

- Topics are defined as a weighted sum of unique words
- Documents are defined as a weighted sum of topics

Popular methods:

Latent Dirichlet Allocation (LDA)

Latent Semantic Analysis (LSA)

Non-negative Matrix Factorization (NMF)

Application of Topic Modeling

Motivation: Uncover the main themes in Cyclone Amphan discourse

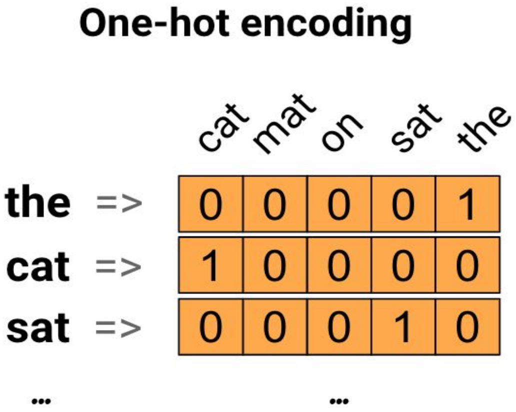
Example 1: Track Themes in Discourse Across Verified and Unverified Users

- We can determine the different themes between public figures and individuals

Example 2: Use dynamic topic models to show changes in discourse over time

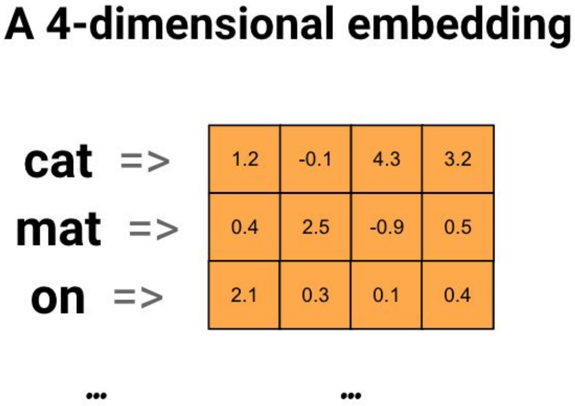
- Determine when topics develop and changes
- Analysis by region could help identify local trends in discourse
- Identify themes that explain the situation

Word Embeddings



One-hot encoding

		cat	mat	on	sat	the
the	=>	0	0	0	0	1
cat	=>	1	0	0	0	0
sat	=>	0	0	0	1	0
…		…				



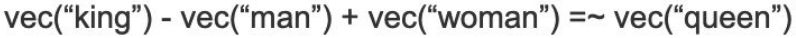
A 4-dimensional embedding

cat	=>	1.2	-0.1	4.3	3.2
mat	=>	0.4	2.5	-0.9	0.5
on	=>	2.1	0.3	0.1	0.4
…		…			

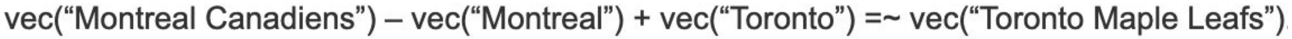
Machine Learning Models take vectors as inputs. When working with text data, it is necessary to convert these strings into numbers before it is fed to a model.

“Word embeddings give us a way to use an efficient, dense representation in which similar words have a similar encoding.”

Application of Word Embeddings



vec(“king”) - vec(“man”) + vec(“woman”) =~ vec(“queen”)



vec(“Montreal Canadiens”) – vec(“Montreal”) + vec(“Toronto”) =~ vec(“Toronto Maple Leafs”).

Motivation: Create dense representations of words and generate a spatial relationship among them. This is a core method that allows the application many other techniques.

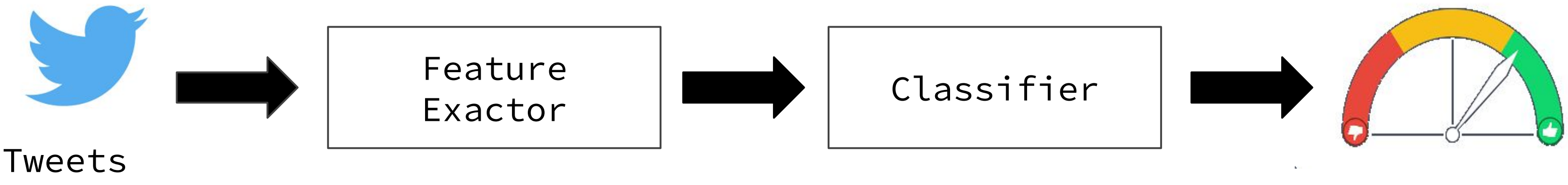
Example 1: Understanding the sentimental value of a tweet. The application of clustering techniques allow the extraction of the associated sentiment for each tweet.

Example 2: Data visualization. Tweets and words within tweets be visualized through the projection of these vectors into a 2d or 3d space, containing useful information.

Popular methods:

- Bag-of-words
- Skip-gram (Word2Vec)
- Doc2Vec
- Continuous bag-of-words (CBOW)

Sentiment Analysis



Tweets

Feature Exactor

Classifier

1. Break each text document down into its component parts (sentences, phrases, tokens and parts of speech)
2. Identify each sentiment-bearing phrase and component
3. Assign a sentiment score to each phrase and component (-1 to +1)
4. Optional: Combine scores for multi-layered sentiment analysis

Popular approaches:

1)Rule Based

2)Automatic

3)Hybrid

Application of Sentiment Analysis

Sentiment Analysis

My experience so far has been fantastic!

The product is ok I guess

Your support team is useless

POSITIVE
NEUTRAL
NEGATIVE
MonkeyLearn

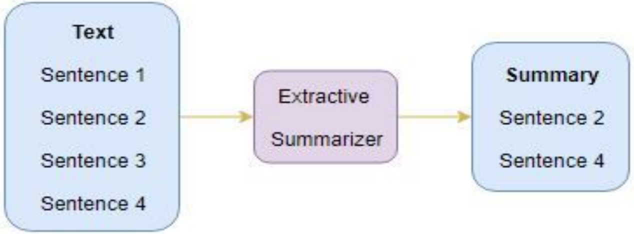
Motivation: Find the attitude of text to be either positive or negative to get an idea of how people feel about certain actions and events

Example 1: Find all tweets having to do with the government response to the cyclone and analyze how people are reacting to it

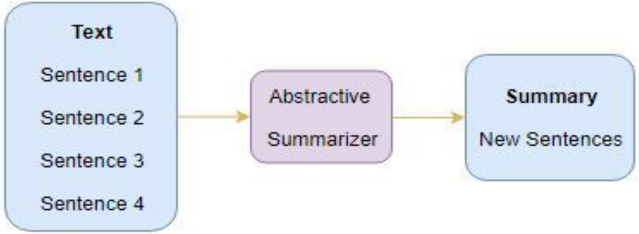
Example 2: Use sentiment analysis to find all people reacting very negatively to see if any insight can be gained on those worse affected by the events happening

Text Summarization

Extractive Summarization



Abstractive Summarization



“**Text summarization** is the task of producing a concise and fluent summary while preserving key information content and overall meaning!”

- **Extractive** - Involves pulling key phrases from the source text and combining them to make a summary. Techniques involve ranking the relevance of phrases in order to choose only those most relevant.
- **Abstractive** - Involves generating entirely new phrases and sentences to capture the meaning of the source. This is a more challenging approach and closer to how humans generate summaries.

Popular methods:

Word Frequency driven approaches

Learning to Rank Methods

Graph Representations

Deep Learning (Attention Models, RNNs)

Application of Text Summarization

Motivation: Summarize the information relevant to a set of themes (**topics**) - critical for post-disaster relief operations, such as Resource requirements and availability, infrastructure damage, restoration, etc.

Example 1: Extract One Tweet (Extractive) or a combination of tweets (Abstractive) to represent a cluster of tweets that are topically relevant or textually similar. **Representative Tweet** - Conveys maximum information by minimizing redundancy and overlap.

Example 2: Track updates on specific situations - No.of Casualties, Affected People, etc and use text-based features to understand what makes a tweet important or relevant to a topic.