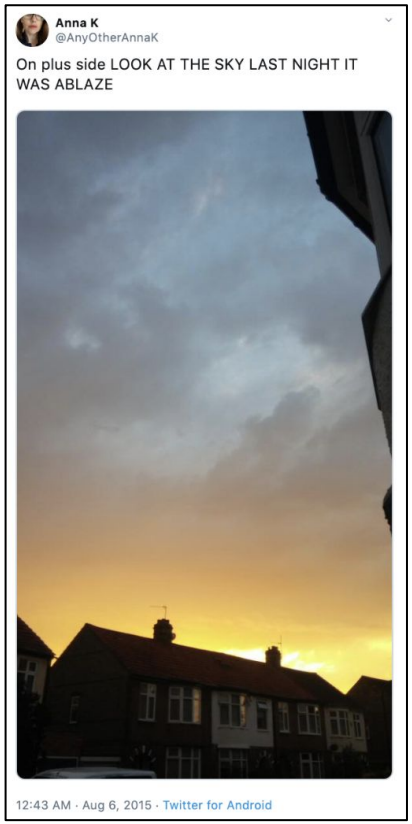
Natural Language Processing with Disaster Tweets

Research Project
CS 834: Introduction to Information Retrieval

Presented by Tarannum Zaki

Department of Computer Science
Old Dominion University, Norfolk, Virginia
Sep 19, 2023

Is the author's Tweet actually announcing a real disaster?



https://www.kaggle.com/competitions/nlp-getting-started

Research Questions

- Is the Tweet about a real disaster or not?
- How effectively the prediction is performed?

Aim of the Project

To build a machine learning model that predicts whether Tweets are about real disasters or not using information retrieval methods.

Proposed Method : TF-IDF

Term Frequency(TF) - Inverse Dense Frequency(IDF)

- A score to highlight each word's relevance in the entire document.
- Helps machine to read words in numbers.

TF = (Number of repetitions of word in a document) / (# of words in a document)

IDF =Log[(# Number of documents) / (Number of documents containing the word)]

TF-IDF Score = TF * IDF

https://medium.com/analytics-vidhya/tf-idf-term-frequency-technique-easiest-explanation-for-text-classification-in-nlp-with-code-8ca3912e58c3

TF-IDF : Advantages

- How useful a word is to a sentence?
- How useful a word is to a document?
- Helps ignore misspelled words.

Document 1 It is going to rain today.

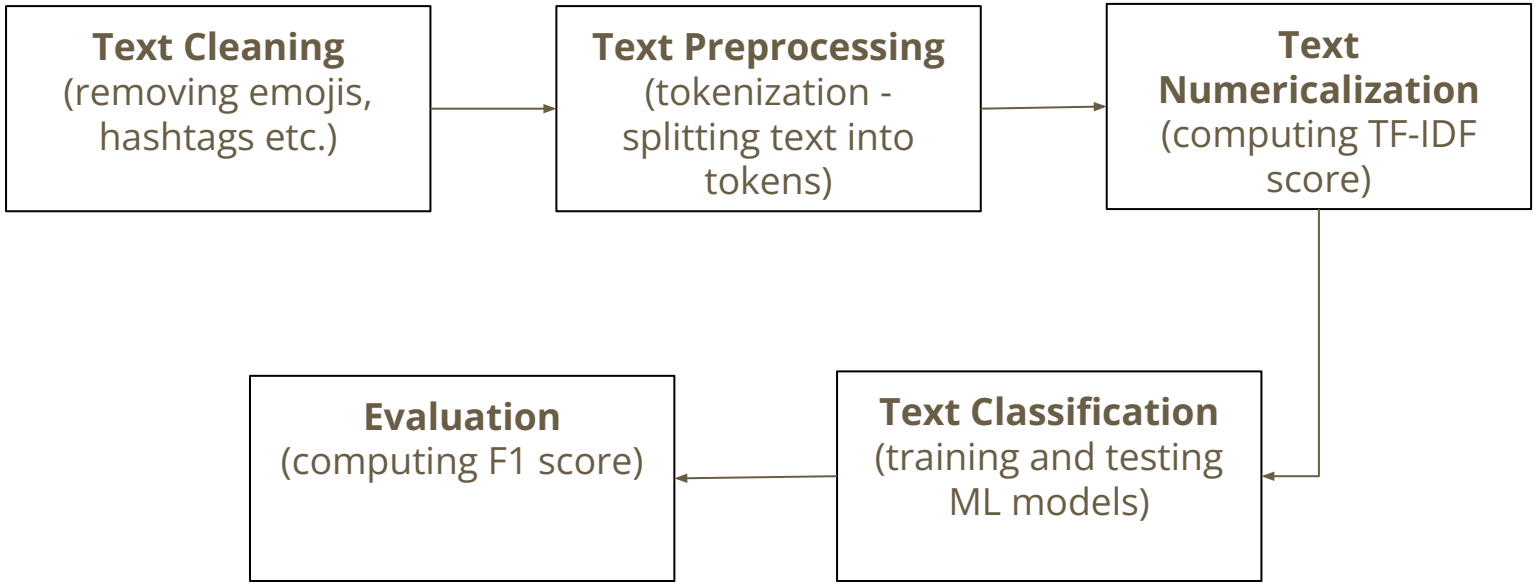
Document 2 Today I am not going outside.

Document 3 I am going to watch the season premiere.

Words/ Documents	going	to	today	i	am	it	is	rain
Document 1	0	0.07	0.07	0	0	0.17	0.17	0.17
Document 2	0	0	0.07	0.07	0.07	0	0	0
Document 3	0	0.05	0	0.05	0.05	0	0	0

https://medium.com/analytics-vidhya/tf-idf-term-frequency-technique-easiest-explanation-for-text-classification-in-nlp-with-code-8ca3912e58c3

Framework: Tasks



Data

- A data set of 10,000 Tweets that were hand classified.
- Split into training and testing data set.
- Data fields:
  - ID
  - **Tweet text**
  - Keyword
  - Location
  - Target

id	keyword	location	text	target
50	ablaze	AFRICA	#AFRICANBAZE: Breaking news:Nigeria flag set ablaze in Aba. http://t.co/2nndBGwyEi	1
52	ablaze	Philadelphia, PA	Crying out for more! Set me ablaze	0

https://www.kaggle.com/competitions/nlp-getting-started/data

Evaluation

Calculate F1 score between the predicted and expected answers:

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

https://www.kaggle.com/competitions/nlp-getting-started

Risks & Mitigations

- TF-IDF does not capture the semantic relationships between words.
- TF-IDF doesn't take into account the surrounding context of terms.
- TF-IDF heavily relies on term frequencies.

Hardware/Software Requirements

Hardware: Personal computer

Software: Jupyter Notebooks (Kaggle), Python