Text Analytics
Natural Language Processing

Important Resource - https://web.stanford.edu/~jurafsky/slp3/ed3bookfeb3_2024.pdf

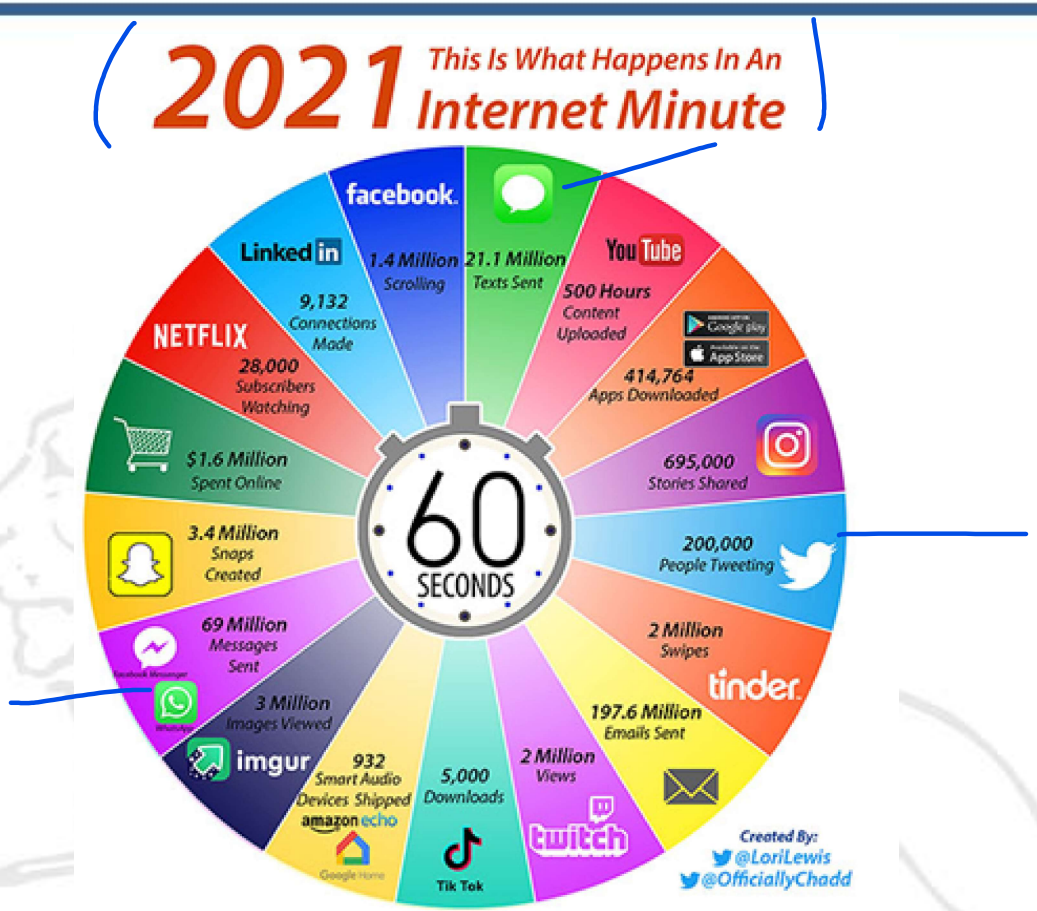
Sumit Kumar Yadav

Department of Management Studies

https://web.stanford.edu/~jurafsky/

Sunday 9th June, 2024

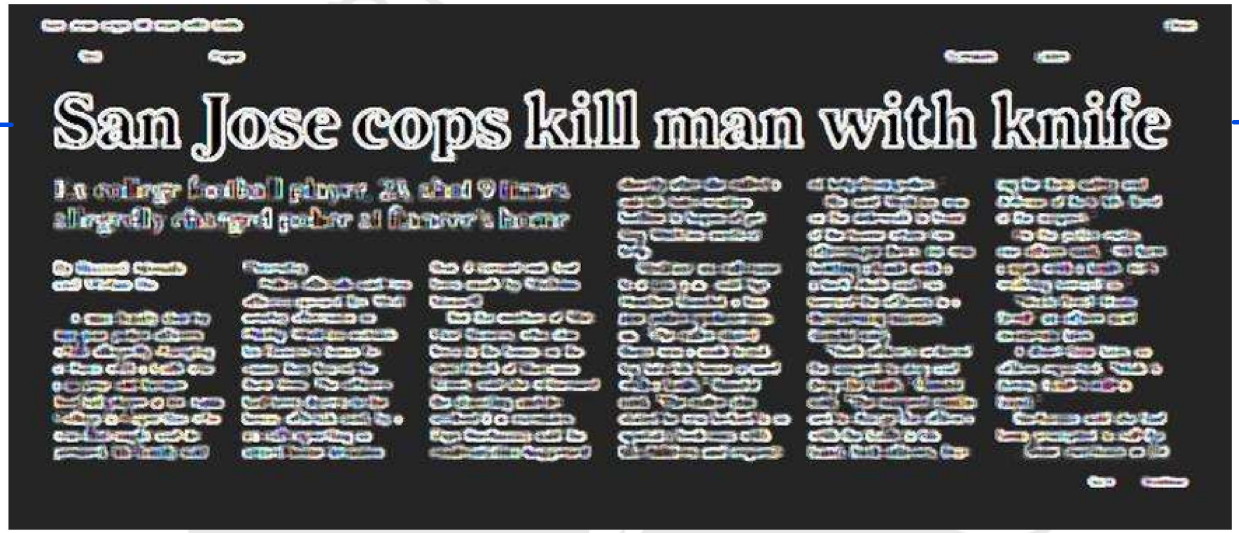
Why is Text Analytics Important?



- In the digital age, vast amounts of text data are generated daily from social media, websites, emails, and documents.
- Text analytics helps in converting this unstructured data into structured data that can be analyzed.

Why is NLP hard

- Who does ‘he’ refer to in the following?
  - Rahim helped Pooja. He was kind.
  - Rahim helped Mohan. He was kind.
  - Rahim helped Kiran. He was kind....
- Who had the knife?



Challenges in Text Analytics

- **Ambiguity:** Natural language is inherently ambiguous. Words can have multiple meanings based on the context.
- **Sarcasm and Irony:** Detecting sarcasm and irony in text is challenging but crucial for accurate sentiment analysis.
- **Language Diversity:** Multiple languages and dialects increase the complexity of text analysis.
- **Volume and Velocity:** The sheer volume of text data and the speed at which it is generated pose significant challenges in terms of processing and analysis.

Key Techniques in Text Analytics

1. **Tokenization:** Breaking down text into individual words or phrases.
2. **Stop Words Removal:** Eliminating common words that do not add much meaning to the text.
3. **Stemming and Lemmatization:** Reducing words to their base or root form.
4. **Term Frequency-Inverse Document Frequency (TF-IDF):** Identifying the importance of words in a document relative to a collection of documents.
5. **Sentiment Analysis:** Determining the sentiment expressed in a piece of text.

Applications of Text Analytics

- **Sentiment Analysis:** Analyzing customer feedback to determine overall sentiment about a product or service.
- **Topic Modeling:** Discovering the underlying themes or topics in a large corpus of text.
- **Summarization:** Automatically generating a concise and coherent summary of a large text.
- **Named Entity Recognition (NER):** Identifying and classifying key elements in text into predefined categories, such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

Conclusion

- Text analytics is a powerful tool that helps in transforming unstructured text data into actionable insights.
- Despite its challenges, advancements in AI and machine learning are continually improving the accuracy and efficiency of text analytics processes.
- As technology advances, the scope and application of text analytics will continue to expand, offering even greater insights into the vast quantities of text data produced every day.

Text Analytics

What is Text?

- Text is the data present in the form of natural language
- The data captured from the native natural language forms a collection of text corpuses
- Text data can be represented as :
  - Readable documents
  - Audio files
  - Image

Case Study - Lenovo K8 Phone Reviews from Amazon

(1.) Frequency based Analysis

- Bar-graph
- Word-cloud

1.) NLTK (Natural Language Tool Kit)

2.) Re (Regular Expressions)

- 14000+ reviews of text
- Tokenization
- Stemming
- Lemmatization

(Jurafsky)

(1.) Dealing with negation

(2) Stop words

What is coming up?

- Sentiment Analysis
- Converting text to numbers
- DTM and TF-IDF matrix
- Some more text cleaning exercises and examples
- Cosine Similarity