



Natural Language Processing

NLP Pipeline

Instructor: Moushmi Dasgupta

Connect up with me on LinkedIn

Some of the material is from Georgia Institute of Technology, Atlanta, GA, USA.

AGENDA

Module 2

NLP Pipeline

2.1 NLP Pipeline

2.2 Data Acquisition

2.3 Text Extraction and Cleanup

2.4 Text Representation

2.5 Model Deployment and Monitoring

Pre-requisite:

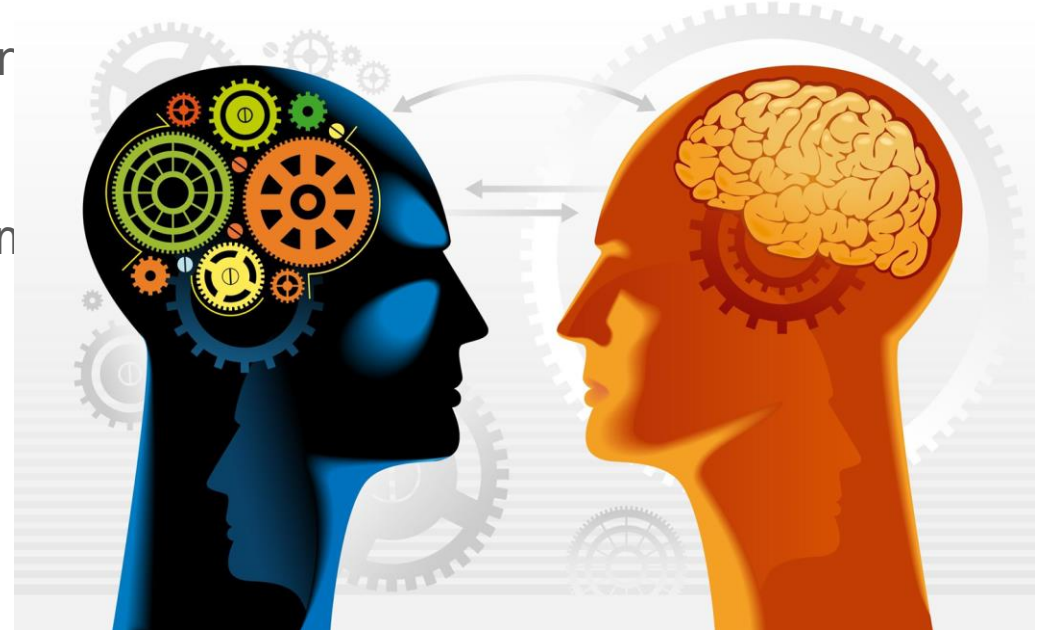
1. Python programming
2. An understanding of Machine Learning
3. Fundamentals of NLP Understanding and hands-on with Python programming on and IDE
3. Invest in attending classroom sessions (Weekly 1 or 2 classes of 3+ hours duration)
4. Invest in yourself with 1 hour of self study everyday



Natural Language Processing

Natural Language Processing

1. Natural Language Processing is a subfield of artificial intelligence concerned with methods of communication between computers and natural languages such as english, hindi, etc.
2. Objective of Natural Language processing is to perform useful tasks involving human languages like
 - Sentiment Analysis
 - Machine Translation
 - Part of Speech Tags
 - Human-Machine communication(chatbots)



Why study NLP?

- ❑ Language is involved in most of the activities that involve interaction between humans, e.g. reading, writing, speaking, listening.
- ❑ Voice can be used as an interface for interactions between humans and machines e.g. cortana, google assistant, siri, amazon alexa.
- ❑ There is massive amount of data available in text format which can be used to derive insights from using NLP, e.g. blogs, research articles, consumer reviews, literature, discussion forums.



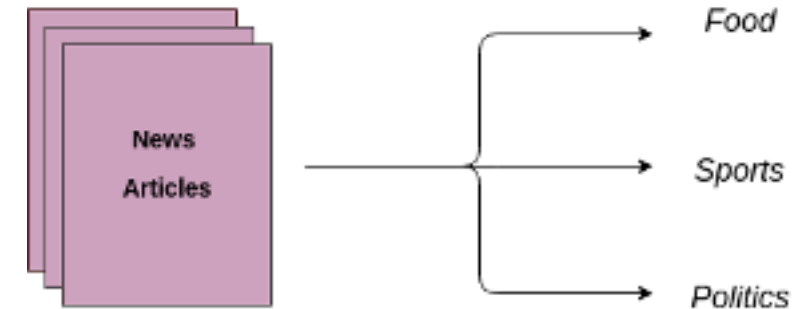
Different Tasks in NLP

- **Text Classification**

- Sentiment Analysis: Determining the general context of a review, whether it is positive or negative or neutral.
- Consumer Complaints Classification: Categorizing complaints on consumer forums to respective departments.

- **Machine Translation**

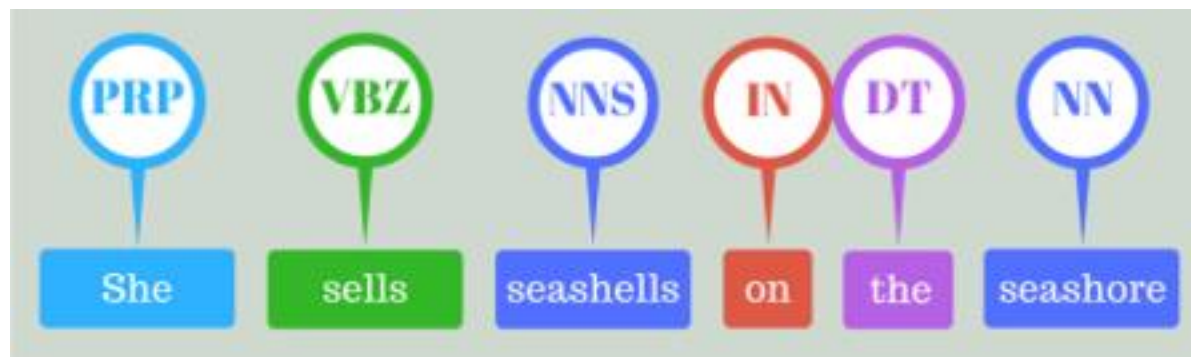
- Improving human-human interaction by translating sentences from one language to another.



Different Tasks in NLP

- **Part of Speech Tagging**

- In corpus linguistics, part-of-speech tagging (POS tagging or PoS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context.
- A simplified form of this is the identification of words as nouns, verbs, adjectives, adverbs, etc.
- Tagset: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html



Movie Ratings

positive

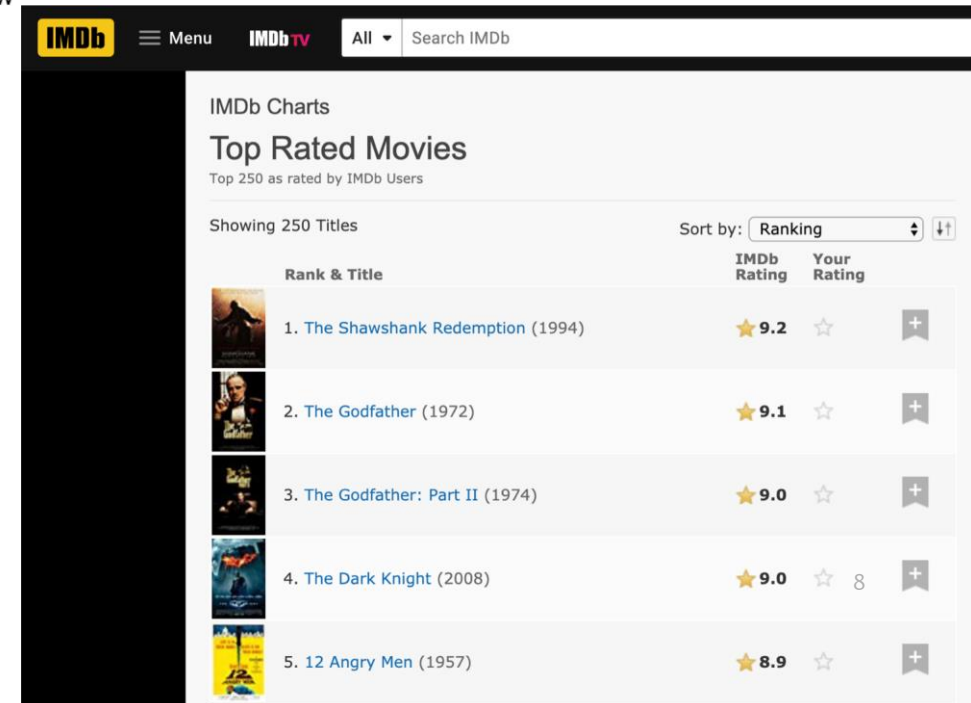
“... is a film which still causes real, not figurative, chills to run along my spine, and it is certainly the bravest and most ambitious fruit of Coppola's genius”

Roger Ebert, Apocalypse Now

- “I hated this movie. Hated hated hated hated hated this movie. Hated it. Hated every simpering stupid vacant audience-insulting moment of it. Hated the sensibility that thought anyone would like it.”

Roger Ebert, North

negative



The screenshot shows the IMDb website's 'Top Rated Movies' chart. The header includes the IMDb logo, a menu icon, 'IMDb TV', and a search bar. The main content area is titled 'IMDb Charts' and 'Top Rated Movies', with a subtitle 'Top 250 as rated by IMDb Users'. It indicates 'Showing 250 Titles' and a 'Sort by: Ranking' dropdown. The chart lists the top 5 movies with their rank, title, year, IMDb rating (stars), and a 'Your Rating' column with a star icon and a plus button.

Rank & Title	IMDb Rating	Your Rating
1. The Shawshank Redemption (1994)	★ 9.2	☆ +
2. The Godfather (1972)	★ 9.1	☆ +
3. The Godfather: Part II (1974)	★ 9.0	☆ +
4. The Dark Knight (2008)	★ 9.0	☆ 8 +
5. 12 Angry Men (1957)	★ 8.9	☆ +

Customer Review



NOT DISHWASHER SAFE

Reviewed in the United States on April 5, 2019

Color: Blue | **Verified Purchase**

Used the bottle for one day. There was a slight lid leak, but I was willing to overlook that because I liked the other aspects of the product. Put it in the dishwasher with my other water bottles, air dry, and it melted. There is nothing in the product description that indicates it is not dishwasher safe, nor was there a product sheet included with the bottle indicating to hand wash only. I have a number of plastic water bottles that I routinely send through the dishwasher on this setting and have never had a problem. Extremely disappointed!

19 people found this helpful

Helpful

Comment

Report abuse



Makes Drinking Water Fun

Reviewed in the United States on March 31, 2019

Color: Transparent | **Verified Purchase**

It is always a challenge to drink the recommended amount of water each day, so important for health. This bottle makes it fun while serving as a reminder to keep drinking! Bottle is good quality, handle makes it easy to lift.



14 people found this helpful

Customer reviews



4.5 out of 5

451 customer ratings



By feature

Sturdiness	★★★★☆ 4.5
Flavor	★★★★☆ 4.5
Durability	★★★★☆ 4.4

Political Opinion Mining



emilia @PoliticalEmilia · 43m

As somebody whose immediate family are **immigrants** from Iran, I want to remind that this isn't the fault of Iranian Americans. Most of us want no more war in the Middle East.

Take your anger out at your government leaders, not at us. We have nothing to do with it. [#IranAttacks](#)

81

239

1.9K



Nithya Raman @nithyavraman · Jan 6

LA is one of the most **immigrant**-rich cities in the US.

Almost 50% of residents are foreign-born. 10% are undocumented.

As Trump works to implement his racist agenda, what are our elected officials doing to defend **immigrant** Angelenos?

The answer: infuriatingly little. (thread)

55

138

606



Brigitte Gabriel @ACTBrigitte · 3m

Thank Goodness there were ZERO U.S. casualties from the attacks Iran made tonight.

President **Trump** is monitoring the situation with his top leaders right now.

I've never felt more comfortable with a leader at the helm, than I do tonight with President **Trump** in office.

21

145

413



Palmer Report @PalmerReport · 1m

So a foreign nation fired missiles at U.S. troops tonight, and the President of the United States ISN'T addressing the nation? How far gone is Donald **Trump**? His handlers don't even trust him to read a speech off a teleprompter anymore.

15

74

225



Andrea Chalupa @AndreaChalupa · 7m

Trump is betting on Iran doing something so horrific to Americans that we rally around the flag, and the 2020 election becomes a mindless debate of who's "patriotic" vs. who's anti-war ("weak" on Iran).

47

147

425



Female or Male Author?

1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochinchina; the central area with its imperial capital at Hue was the protectorate of Annam...
2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets...

S. Argamon, M. Koppel, J. Fine, A. R. Shimoni, 2003. "Gender, Genre, and Writing Style in Formal Written Texts," *Text*, volume 23, number 3, pp. 321–346

Is This Spam?

Subject: Important notice!

From: Stanford University <newsforum@stanford.edu>

Date: October 28, 2011 12:34:16 PM PDT

To: undisclosed-recipients;;

Greats News!

You can now access the latest news by using the link below to login to Stanford University News Forum.

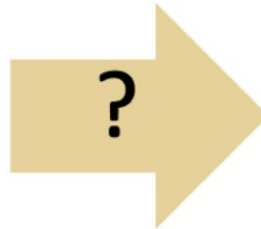
<http://www.123contactform.com/contact-form-StanfordNew1-236335.html>

Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information about the new services.

© Stanford University. All Rights Reserved.

What Is the Subject of This Article?

MEDLINE Article



MeSH Subject Category Hierarchy

- Antagonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- ...

This Class

- Basic representations of text data for classification
- Three linear classifiers
 - Naïve Bayes
 - Perception
 - Logistic regression

Some Direct Text Classification Applications

Task	x	y
Language identification	text	{English, Mandarin, Greek, ...}
Authorship attribution	text	{jk rowling, james joyce, ...}
Sentiment classification	text	{positive, negative, neutral, mixed}

Linear Classification on the Bag of Words

- Let $f(x, y)$ score the compatibility of bag-of-words x and label y , then

$$\hat{y} = \operatorname{argmax}_y f(x, y)$$

- In a **linear classifier**, this scoring function has a simple form:

$$f(x, y) = \mathbf{w} \cdot \mathbf{f}(x, y) = \sum_{i=1}^n w_i \cdot f_i(x, y)$$

- where \mathbf{w} is a vector of weights, and f is a **feature function**

Summary of Linear Classification

	Pros	Cons
Naive Bayes	Simple, probabilistic, fast Closed-form	Not very accurate
Logistic Regression	Error-driven learning, regularized	More difficult to implement

Different Tasks in NLP

- **Word Segmentation**
 - In some languages, there is no space between words, or a word may contain smaller syllables. In such languages, word segmentation is the first step of NLP systems.
- **Semantic Analysis**
 - Semantic analysis of a corpus (a large and structured set of texts) is the task of building structures that approximate concepts from a large set of documents.
 - Application of Semantic Analysis:
 - Text Similarity
 - Context Recognition
 - Sentence Parsing
 - Topic Modelling

Why NLP is hard?

- ❑ Languages are changing everyday, new words, new rules, etc.
- ❑ The number of tokens is not fixed. A natural language can have hundreds of thousands of different words, new words are created on the fly.
- ❑ Words can have different meanings depending on context, and they can acquire new meanings over time (apple(a fruit), Apple(the company)], they can even change their parts of speech(Google --> to google).
- ❑ Every language has its own uniqueness.
- ❑ Like in the case of English we have words, sentences, paragraphs and so on to limit our language. But in Thai, there is no concept of sentences.



Pre-processing Steps

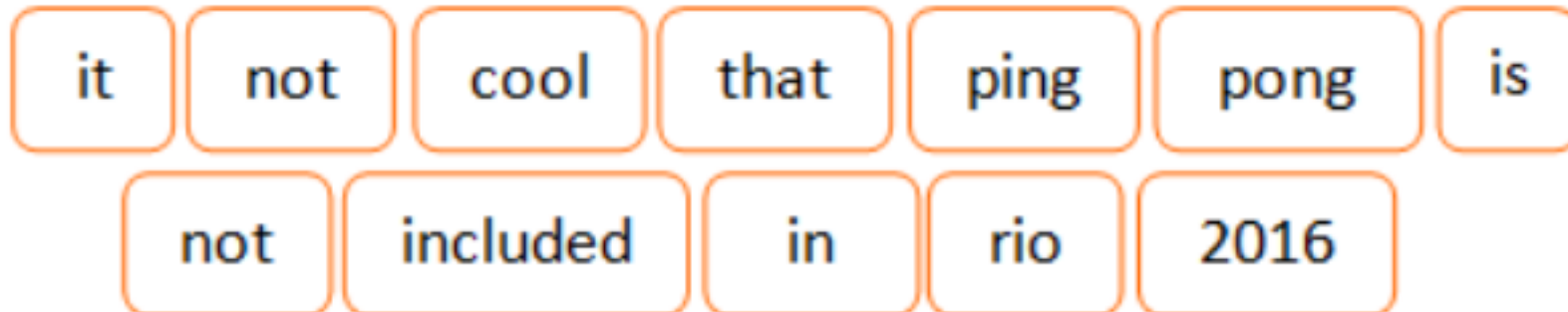
Tokenization

- Tokenization is the task of taking a text or set of text and breaking it up into its individual tokens.
- Tokens are usually individual words (at least in languages like English).
- Tokenization can be achieved using different methods. Most common method is Whitespace tokenizer and Regexp Tokenizer. We will use them in our case study.

it not cool that ping pong is not included in rio 2016

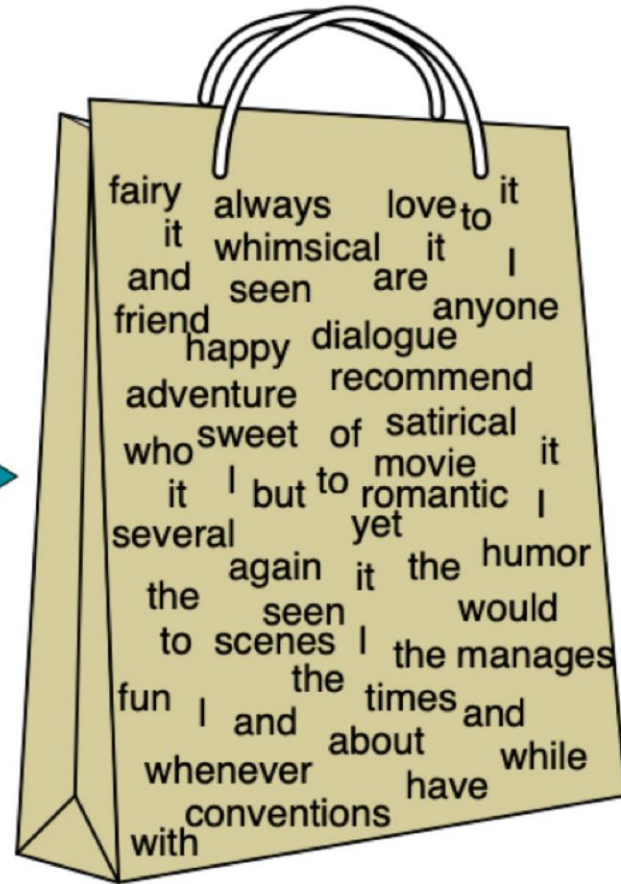


Tokenization



Bag-of-Words

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

The Bag-of-Words

- One challenge is that the sequential representation (w_1, w_2, \dots, w_T) may have a different length T for every document.

- The bag-of-words is a fixed-length representation, which consists of a vector of word counts:

\mathbf{w} = It was the best of times, it was the worst of times

$\mathbf{x} = [\underbrace{\text{aardvark}}_0, \dots, \underbrace{\text{best}}_1, \dots, \underbrace{\text{it}}_2, \dots, \underbrace{\text{of}}_2, \dots, \underbrace{\text{zyther}}_0]$

- The length of \mathbf{x} is equal to the size of the vocabulary V
- For each \mathbf{x} , there may be many possible \mathbf{w} , depending on word order.

Stop Words Removal

- Stopwords are common words that carry less important meaning than keywords.
- When using some bag of words based methods, i.e, countVectorizer or tfidf that works on counts and frequency of the words, removing stopwords is great as it lowers the dimensional space.
- Not always a good idea?
 - When working on problems where contextual information is important like machine translation, removing stop words is not recommended.

```
> stopwords("english")
[1] "i"      "me"      "my"      "myself"  "we"
[6] "our"    "ours"    "ourselves" "you"     "your"
[11] "yours"  "yourself" "yourselves" "he"      "him"
[16] "his"    "himself" "she"      "her"     "hers"
[21] "herself" "it"      "its"      "itself"  "they"
[26] "them"   "their"   "theirs"   "themselves" "what"
[31] "which"  "who"     "whom"    "this"    "that"
[36] "these"  "those"   "am"      "is"      "are"
[41] "was"    "were"    "be"      "been"    "being"
[46] "have"   "has"     "had"     "having"  "do"
[51] "does"   "did"     "doing"   "would"   "should"
[56] "could"  "ought"   "i'm"     "you're"  "he's"
[61] "she's"  "it's"    "we're"   "they're" "i've"
[66] "you've" "we've"   "they've" "i'd"     "you'd"
[71] "he'd"   "she'd"   "we'd"    "they'd"  "i'll"
[76] "you'll" "he'll"   "she'll"  "we'll"   "they'll"
[81] "isn't"  "aren't"  "wasn't"  "weren't" "hasn't"
[86] "haven't" "hadn't"  "doesn't" "don't"   "didn't"
[91] "won't"  "wouldn't" "shan't"  "shouldn't" "can't"
[96] "cannot" "couldn't" "mustn't" "let's"   "that's"
[101] "who's"  "what's"  "here's"  "there's" "when's"
[106] "where's" "why's"   "how's"   "a"       "an"
```


Stemming and Lemmatization

- ❑ The idea of reducing different forms of a word to a core root.
- ❑ Words that are derived from one another can be mapped to a central word or symbol, especially if they have the same core meaning.
- ❑ In stemming, words are reduced to their word stems. A word stem is an equal to or smaller form of the word.
- ❑ “cook,” “cooking,” and “cooked” all are reduced to same stem of “cook.”
- ❑ Lemmatization involves resolving words to their dictionary form. A lemma of a word is its dictionary or canonical form!



Word Features

Bag of Words

- In this model, a text (such as a sentence or a document) is represented as the bag of its words, disregarding grammar and even word order but keeping multiplicity.
- We use the tokenized words for each observation and find out the frequency of each token.
- We define the vocabulary of corpus as all the unique words in the corpus above and below some certain threshold of frequency.
- Each sentence or document is defined by a vector of same dimension as vocabulary containing the frequency of each word of the vocabulary in the sentence.
- The bag-of-words model is commonly used in methods of document classification where the (frequency of) occurrence of each word is used as a feature for training a classifier.

Raw Text

it is a puppy and it
is extremely cute

Bag-of-words vector

it	2
they	0
puppy	1
and	1
cat	0
aardvark	0
cute	1
extremely	1
...	...

Tf-idf Vectors

- Tf-idf (term frequency times inverse document frequency) is a scheme to weight individual tokens.
- One of the advantage of tf-idf is reduce the impact of tokens that occur very frequently, hence offering little to none in terms of information.

*TFIDF score for term i in document $j = TF(i, j) * IDF(i)$*

where

IDF = Inverse Document Frequency

TF = Term Frequency

$$TF(i, j) = \frac{\text{Term } i \text{ frequency in document } j}{\text{Total words in document } j}$$

$$IDF(i) = \log_2 \left(\frac{\text{Total documents}}{\text{documents with term } i} \right)$$

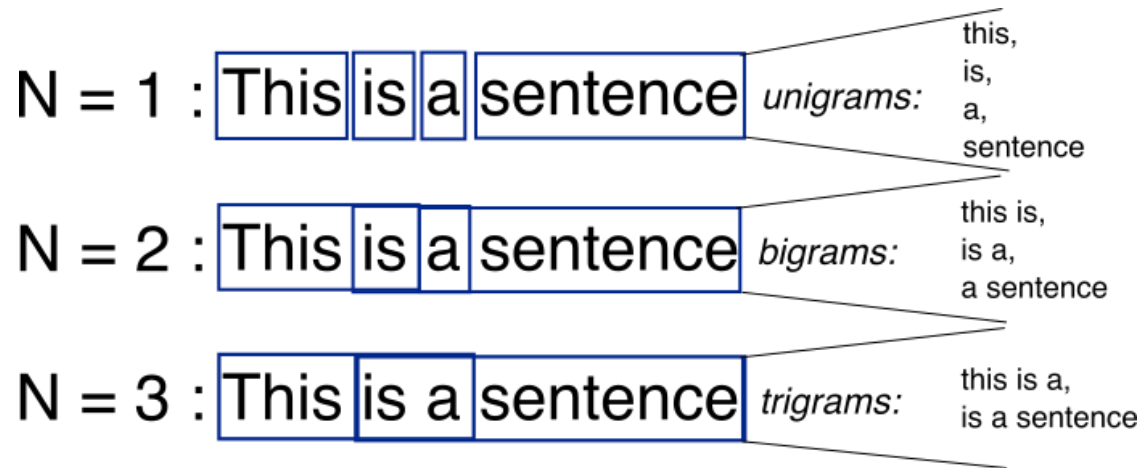
and

t = Term

j = Document

N-gram and Language Model

- Language models are the type of models that assign probabilities to sequence of words.
- N-grams is the most simplest language model. It's a sequence of N-words.
- Bi-gram is a special case of N-grams where we consider only the sequence of two words (Markovian assumption).
- In N-gram models we calculate the probability of Nth words give the sequence of N-1 words. We do this by calculating the relative frequency of the sequence occurring in the text corpus.



Bigram approximation

$$P(w_1^n) = \prod_{k=1}^n P(w_k | w_{k-1})$$

N-gram approximation

$$P(w_1^n) = \prod_{k=1}^n P(w_k | w_{k-N+1}^{k-1})$$

NLP Pipeline Steps



Some of the material is from Georgia Institute of Technology, Atlanta, GA, USA.

Data Acquisition

In the data acquisition step – lots of things to understand.

Data Availability – Where and how much is the Data

Data Acquisition Methods

- From Where
- How Much
- How much is enough
- Clean or not
- Sufficient or not
- Required or not
- Who has the data?
- How to get it?





Text Cleaning / Pre-processing

1. Text Cleaning – HTML tag removing, Spelling checker, etc.
2. Basic Preprocessing —Tokenization(word or sent tokenization, stop word removal, removing digit, lower casing, etc
3. Advance Preprocessing — In this step we do POS tagging, Parsing, etc



Lowercasing

As we know python is case sensitive language.

John, JOHN, john to work the same, converting to lower case is best.

Check out the script from NLTL to convert to lower case.

```
df['text'].str.lower()  
df['text'].apply(lambda  
x:x.lower())
```



What's Next?

🧠 Text Classification