Data ^{X}

Me:

Introduction to Natural Language Processing - I

(Text Based NLP)

By Alexander Fred-Ojala (S2018), Sana Iqbal (F2017)

What is Natural Language Processing (NLP)?

- NLP aims to analyze language and extract meaning.

- Natural Language Processing is a field at intersection of computer science, artificial intelligence, linguistics.

Only 21% of data available in structured form. Most data is unstructred, and textual data is most common.

Why should we care about NLP?

We want computers to understand human language, so that:

1. Computers can communicate with humans.
2. We want computers to perform certain tasks using language data, like:
   a. Text Classification
   b. Information Retrieval
   c. Question Answering
   d. Automatic Captioning / Tagging
   e. Language / Machine Translation
   f. Natural Language generation
   g. Optical character recognition (CNNs)
   h. Document to information pipelines / cleaning up text data

NLP in Industry

1. Search engines (Google, Duckduckgo)
2. Voice Assistants (Alexa, Siri..)
3. Autocompletion
4. Sentiment Analysis
5. Language Translation
6. Chatbots
7. …..

Steps to get textual data into a workable format

1. **CLEAN TEXT DATA FROM NOISE**

   a. Clean text from unnecessary symbols, URLs, stop words (*I, you, are*) etc.

   b. Lexicon Normalization (get word into basic form): Stemming / Lemmatization.

   c. Convert to lowercase, no punctuation etc. Standardizing features.

Steps to get textual data into a workable format

2. Convert to Features

a. **Tokenize:** Group into single words or n-grams. N-grams as features (combination of words)

   "Blue Cheese" Bigram or *"Blue" "Cheese"* Unigram

b. **Syntactic parsing:** Part of Speech Tagging (PoS) / Grammatical tagging

   i. **Word disambiguation** "*please book a flight*" vs "*I'll read a book on the flight*" (Verb vs Noun)

c. **Entity extraction:** Named Entity Recognition

   i. Example *Sentence – Alexander is a person in Berkeley working at UC Berkeley.*

   ii. *Named Entities – ("person" : "Alexander"), ("org" : "UC Berkeley"), ("location" : "Berkeley")*

   iii. In order to classify entities we can e.g. utilize Google maps API for locations, Wikipedia for people.

d. **Topic extraction** with Latent Dirichlet Analysis (developed at Berkeley). Unsupervised BoW model.

Steps to get textual data into a workable format

2. Convert to Features

- **Term Frequency – Inverse Document Frequency (TF – IDF) w_i,j**
  - **Term Frequency (TF)** – TF for a term “tf_i,j” is defined as the count of a term “t_i” in a document “D_j”
  - **Inverse Document Frequency (IDF)** – IDF for a term is defined as logarithm of ratio of total documents available in the corpus (N) and number of documents containing the term T. (df_i)
- <u>**Other Features:**</u>
  - Word Count
  - Sentence Count
  - Punctuation Counts
  - Industry specific word counts

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j

df_i = number of documents containing i

N = total number of documents

Machine Learning Classification in NLP

Examples:

- Email Spam Identification,
- Topic classification of news
- Sentiment classification
- Organization of web pages by search engines.

In order to classify textual data, same as the other classifiers we have seen in the class before. Train step and validaiton step. We just need to have the data in the correct format.

Other ways of finding text similarity:

Levenshtein distance (number of edits to transform one to another),

Cosine similarity for vectors

Phonetic similarity

What makes NLP challenging?

- Natural languages are **ambiguous**, in comparison to clearly defined computations e.g
  - print(1+1) always prints 2
  - I want to eat honey.   I want to eat honey.
  - It’s chill.

- Natural language **assumes contextual information** is known e.g.
  - I ask you to count the number of boys in the room. You get up and count.
  - To a computer I need to define what is a room, what is a boy and what count means!

Components Natural Languages:

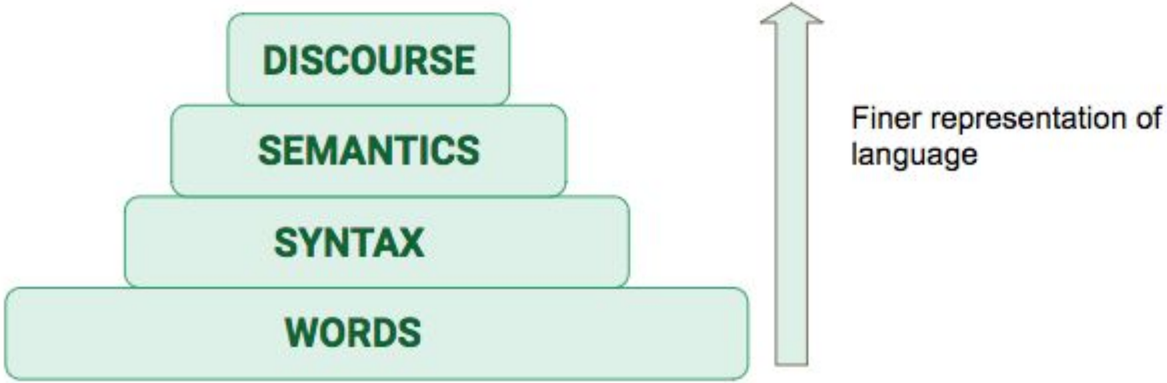
1. **Units:** Words, Sentences, Paragraphs etc.
2. **Syntax:** Parts of Speech, Named Entities, Relations
3. **Semantics:** Lexical (synonyms, definitions) and combination of words
4. **Discourse:** Summarization, categorization of topics

Syntax: The way in which words are put together to form phrases, clauses, or sentences. Grammar related.

Semantics: The study of the meanings of words and phrases in language.

Discourse: The study the meanings of sentences in context with one another.

Different Levels of Analysis



DISCOURSE

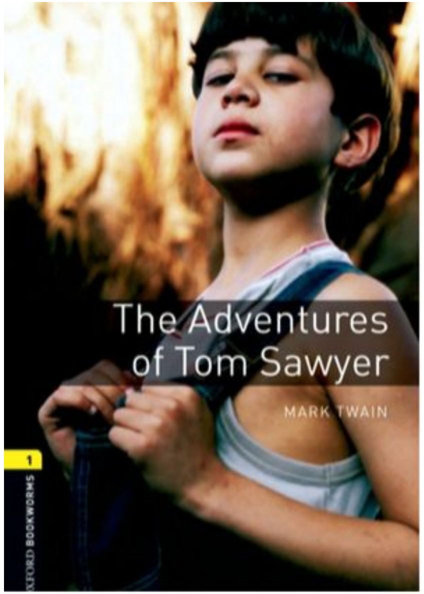
SEMANTICS

SYNTAX

WORDS

Finer representation of language

Consider this text from a novel:



“TOM!”

No answer.

“TOM!”

No answer.

“What is gone with that boy, I wonder? You TOM!”

No answer.

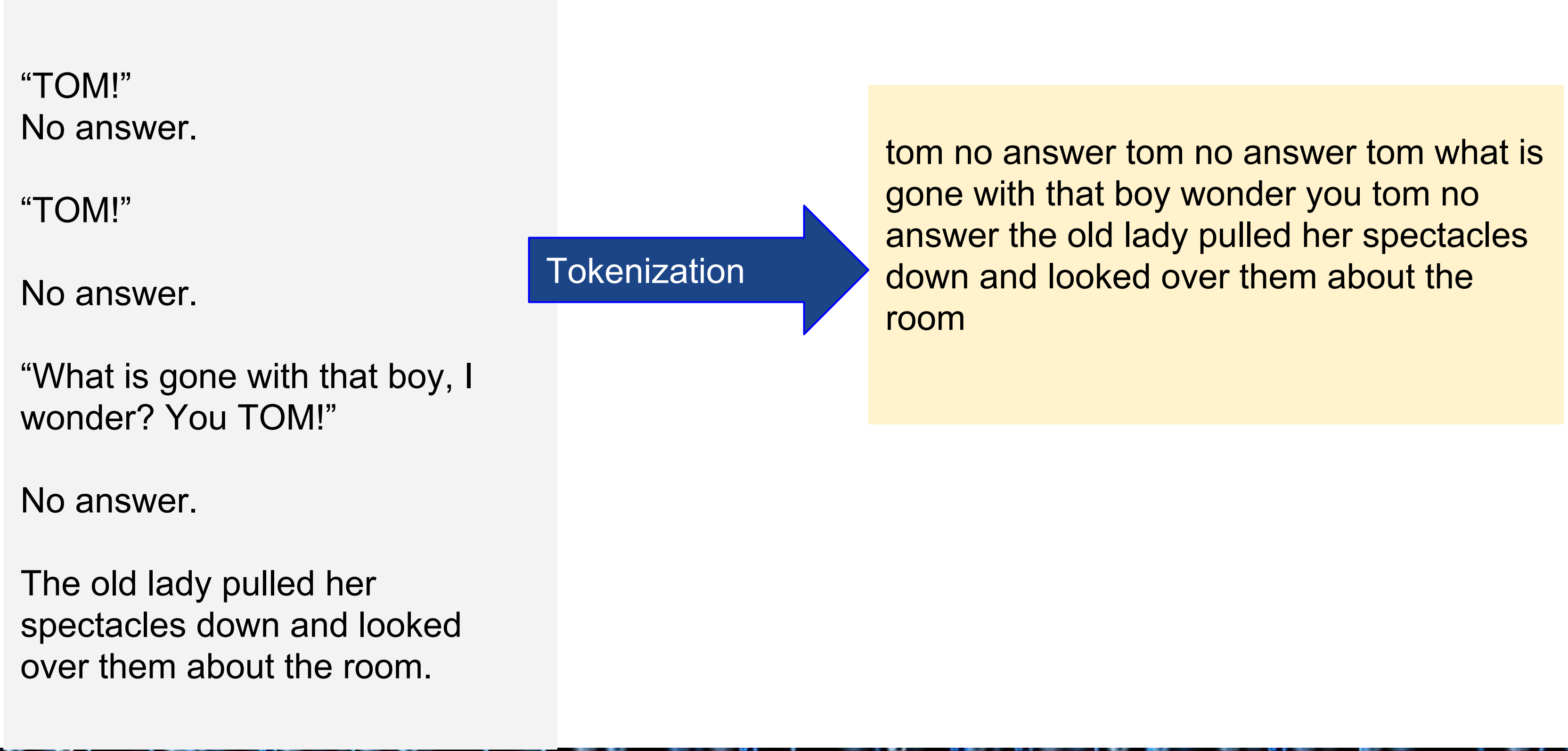
The old lady pulled her spectacles down and looked over them about the room.

Example text from David Bamman material

WORDS

TOKENIZE:

Represent the corpus as a collection of words:



“TOM!”
No answer.

“TOM!”

No answer.

“What is gone with that boy, I wonder? You TOM!”

No answer.

The old lady pulled her spectacles down and looked over them about the room.

tom no answer tom no answer tom what is gone with that boy wonder you tom no answer the old lady pulled her spectacles down and looked over them about the room

Count the word frequencies (BoW features):

tom no answer tom no answer what is gone with that boy wonder you tom no answer the old lady pulled her spectacles down and looked over them about the room

Bag of Words (BOW) representation of this corpus.

Word	count	Word	count
tom	3	old	1
no	3	lady	1
answer	3	pulled	1
what	1	her	1
is	1	spectacles	1
gone	1	down	1
with	1	and	1
that	1	looked	1
boy	1	over	1
wonder	1	them	1
you	1	about	1
the	2	room	1

Bag of Words (BoW) Model of Language

Unigrams (words)

- Every word becomes a feature
- Simplest and most naive way of modeling language
- Oversimplified view of the complex nature of language

Bigrams or trigrams

- Tokens are word pairs or three words
- Ex: “I am hungry now”
  - Bigram: [“I am”, “am hungy”, “hungry now”]
  - Trigram: [“I am hungry”, “am hungry now”]

B. Syntax Analysis:

We want to break down the data to a **structural relationship** between words and how language is constructed.

1. Identify **parts of speech**
   -nouns, verbs,
   adjectives

2. Identify **named entities**
   - look for nouns.

3. Identify **relations or structural phrases.**

PARTS OF SPEECH TAGGING:

--- nouns
--- verbs
--- adjectives

“TOM!”
No answer.

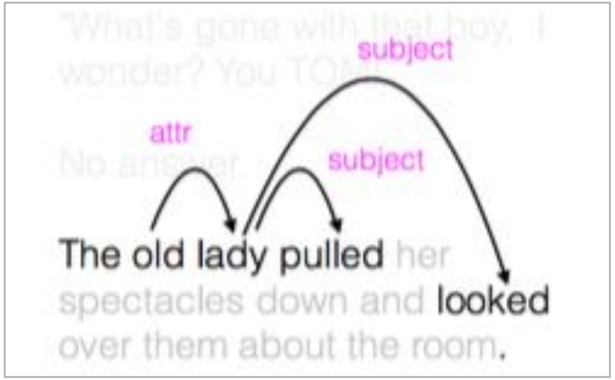
“TOM!”
No answer.

“What is gone with that boy, I wonder? You TOM!”

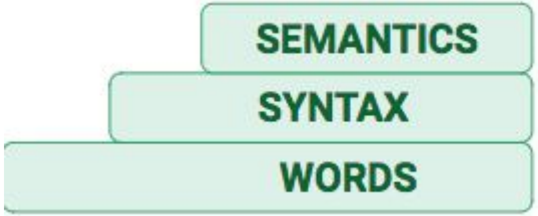
No answer.

The old lady pulled her spectacles down and looked over them about the room.

Chunking:


“What's gone with that boy, I wonder? You TOM!”
subject
attr
No answer.
subject
The old lady pulled her spectacles down and looked over them about the room.

C. Semantic Analysis:



Relationship between different parts of a corpus in creating meaning that is language independent.

Lexical semantics:
Meanings of component words of a corpus
I am good. I am ok. I am well.
Cat is a mammal.

Compositional semantics:
How words or phrases combine to create meaning.

Different semantic frames for a same word:

Apply_heat frame: “Michelle baked the potatoes for 45 minutes .”
Cooking_creation frame: “Michelle baked her mother a cake for birthday.”
Absorb_heat frame: “The potatoes have to bake for more than 30 minutes.”

Not_interesting: Too much theory makes a lecture dry.
Water_evaporated: Your jacket is dry now.

*Some definitions you might come across in NLP:

Lemma: The canonical form of an inflected word, e.g. baked ---->bake
Morphemes: The smallest linguistic unit within a word that can carry a meaning e.g. clueless--->clue, less

D. Discourse Analysis:

Discourse is dialog or communication, which always has an underlying subject and tone. Here we study how sentences in a corpus affect each other and overall meaning of the language.

E.g

"TOM!"
No answer.
"TOM!"
No answer.

"What is gone with that boy, I wonder? You TOM!"

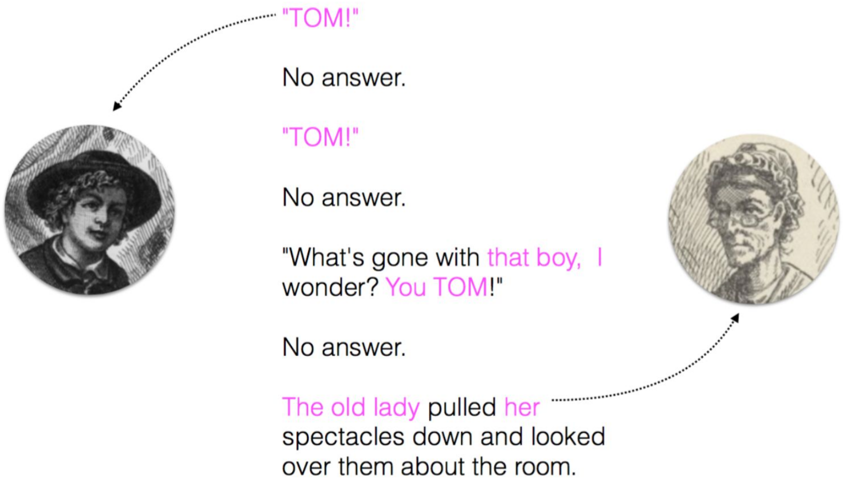
No answer.

The old lady pulled her spectacles down and looked over them about the room.

Reframing

An old woman is looking for a boy named TOM.

Co-references Resolution



"TOM!"

No answer.

"TOM!"

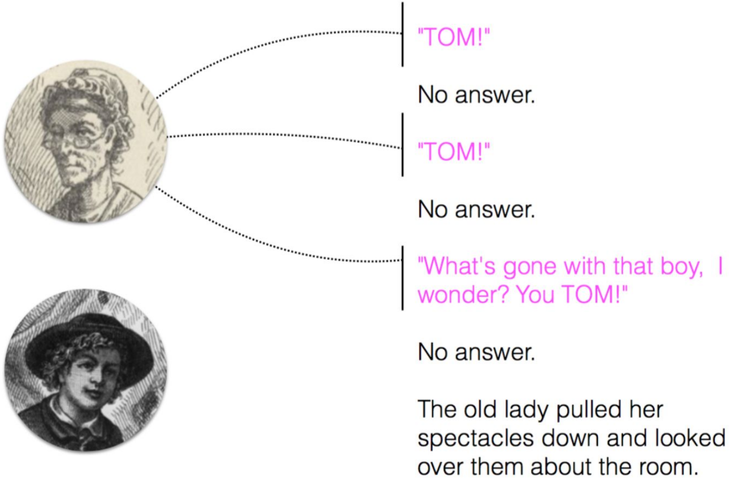
No answer.

"What's gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.

Identify Speakers



"TOM!"

No answer.

"TOM!"

No answer.

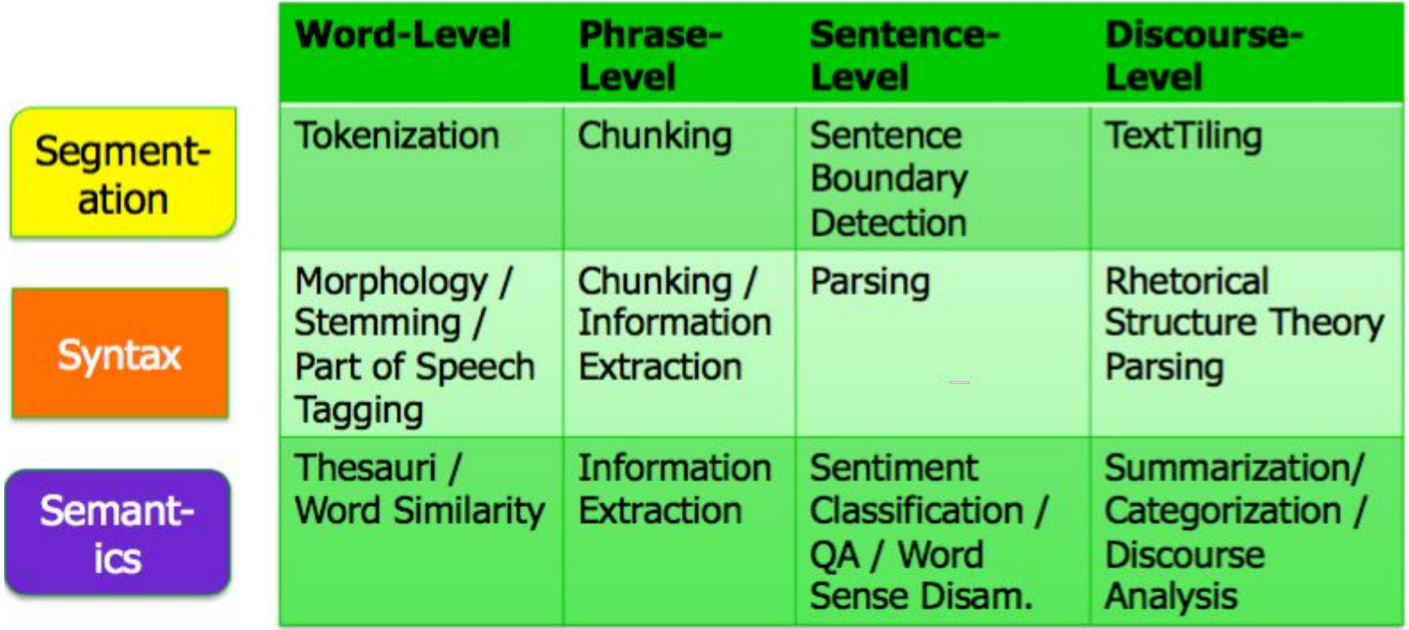
"What's gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.

How is NLP is done? - General Idea

Principles of the traditional NLP



	Word-Level	Phrase-Level	Sentence-Level	Discourse-Level
Segmentation	Tokenization	Chunking	Sentence Boundary Detection	TextTiling
Syntax	Morphology / Stemming / Part of Speech Tagging	Chunking / Information Extraction	Parsing	Rhetorical Structure Theory Parsing
Semantics	Thesauri / Word Similarity	Information Extraction	Sentiment Classification / QA / Word Sense Disam.	Summarization/ Categorization / Discourse Analysis

Table fom Marti Hearst

Natural Language for Machines:

Imagine the scenario, which do you think is more likely?

I want to go outside and take a [________]

I want to go outside and take a walk --- looks most likely Or natural

I want to go outside and take a picture.

I want to go outside and take a stone --- looks least likely

<u>We want our computer to do this job for us!</u>

How can we teach it?

We teach our machine a *LANGUAGE MODEL*

Language Model for Machines:

At a high level:

If we want our machines to do the specific tasks on language , they should be able to understand the language.

To that end we create **language models.**

Language models predict the probability distribution of language expressions given a set of vocabulary.

Language models or the grammar:

Language modeling is the task of estimating likelihood of a **sequence** or **a word** given a sequence.

$P(x1, x2, x3) = P(x1)\ P(x2 \mid x1)\ P(x3 \mid x2, x1)$

EXAMPLE:

P(“sunday is a boring day”) = P(sunday)* P(is|sunday)* P(a|sunday,is)* P(boring|sunday,is,a)* P(day|sunday,is,a,boring)

For longer sentences it becomes very hard to track large dependencies, we make use of **Markov Assumption and MLE of Probability.**

N-grams

The Markov Assumption:

The probability of a future event depends only on a limited history of preceding events.

MLE:

P(wi |w1w2...wi−1)=count(w1...wi)/count(w1...wi-1)

An **n-gram model** is a statistical model of language in which the **previous n−1** words are used to predict the next word.

Unigram Model

- **Likelihood of a word is not dependent on the context of the word**
- Just multiply the probability of each word to get the probability of a sentence.
- $P(w1w2 \ldots wn) \approx \prod P(wi)$

EXAMPLE:

Corpus / Sentence : sunday is a boring day

P(“sunday is a boring day”) = P(sunday)* P(is)* P(a)* P(boring)* P(day)

P(“day|sunday is a boring”) = P(day) = 1/5

Bigram Model

- **Likelihood of a word is dependent on one preceding contextual word:**

  P(wi |w1w2...wi–1)≈P(wi |wi–1)

- **EXAMPLE:**

  P(“day|sunday is a boring ”)   = P(day|boring)

  = Count(“boring day”)/Count(“boring”) = 1

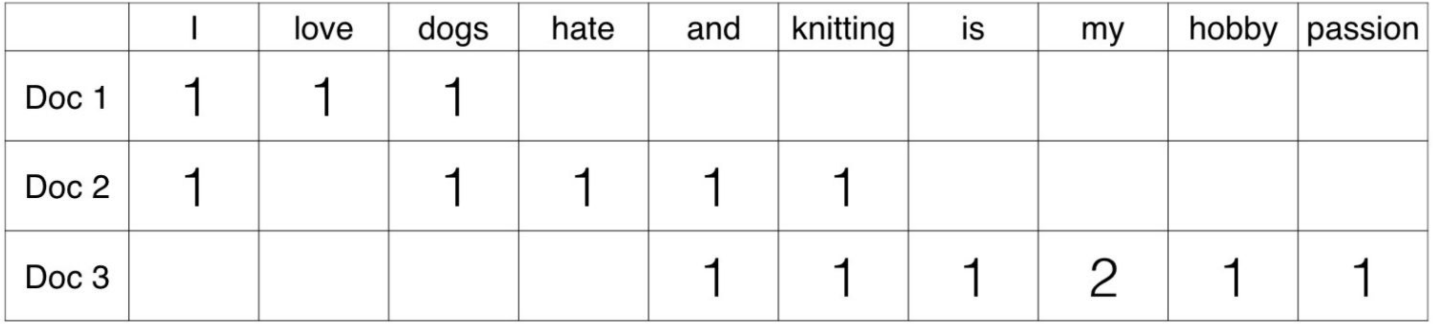
Unigram Bag of Word Feature Matrix

Doc1: I love dogs.

Doc2: I hate dogs and knitting.

Doc3: Knitting is my hobby and my passion

Bag of words:



	I	love	dogs	hate	and	knitting	is	my	hobby	passion
Doc 1	1	1	1							
Doc 2	1		1	1	1	1				
Doc 3					1	1	1	2	1	1

Information Extraction

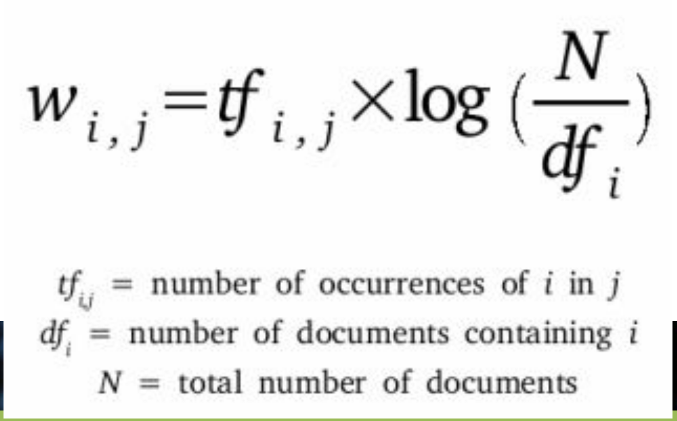
If we want the most relevant document to be delivered back we need to find the document that has our query as the signature word.

IDF: Inverse Document Frequency, which measures how rare a term is in the the documents.

IDF(term(t)) = log_e(Total number of documents / Number of documents with term t in it).

Instead of using, counts we use

tf-idf weights of terms as features.



$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j

df_i = number of documents containing i

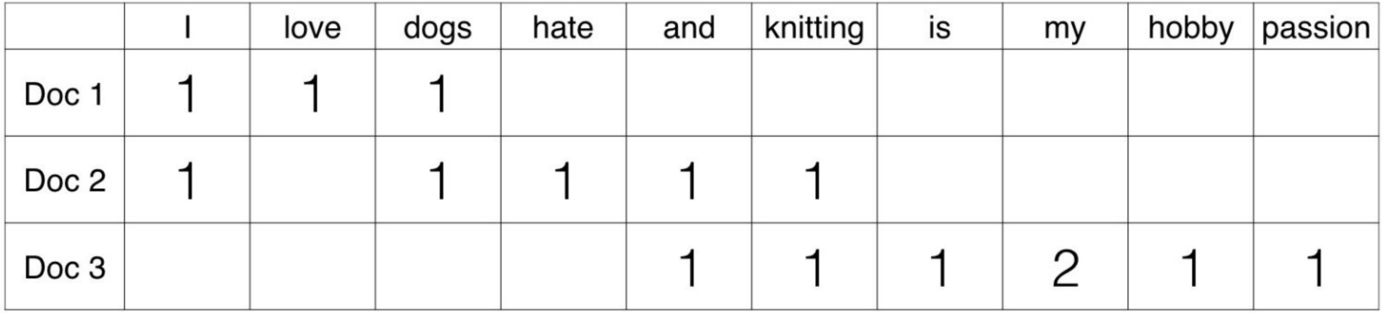
N = total number of documents

Doc1: I love dogs.

Doc2: I hate dogs and knitting.

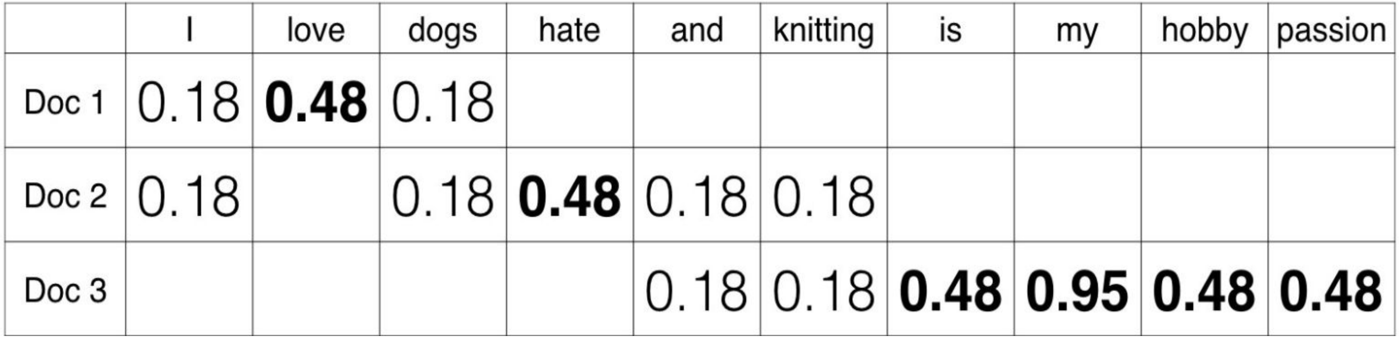
Doc3: Knitting is my hobby d my passion

Counts / Term Frequency(tf):



	I	love	dogs	hate	and	knitting	is	my	hobby	passion
Doc 1	1	1	1							
Doc 2	1		1	1	1	1				
Doc 3					1	1	1	2	1	1

Tf-idf weights:



	I	love	dogs	hate	and	knitting	is	my	hobby	passion
Doc 1	0.18	**0.48**	0.18							
Doc 2	0.18		0.18	**0.48**	0.18	0.18				
Doc 3					0.18	0.18	**0.48**	**0.95**	**0.48**	**0.48**

End of Section

How we can do NLP:

At high level we can have:

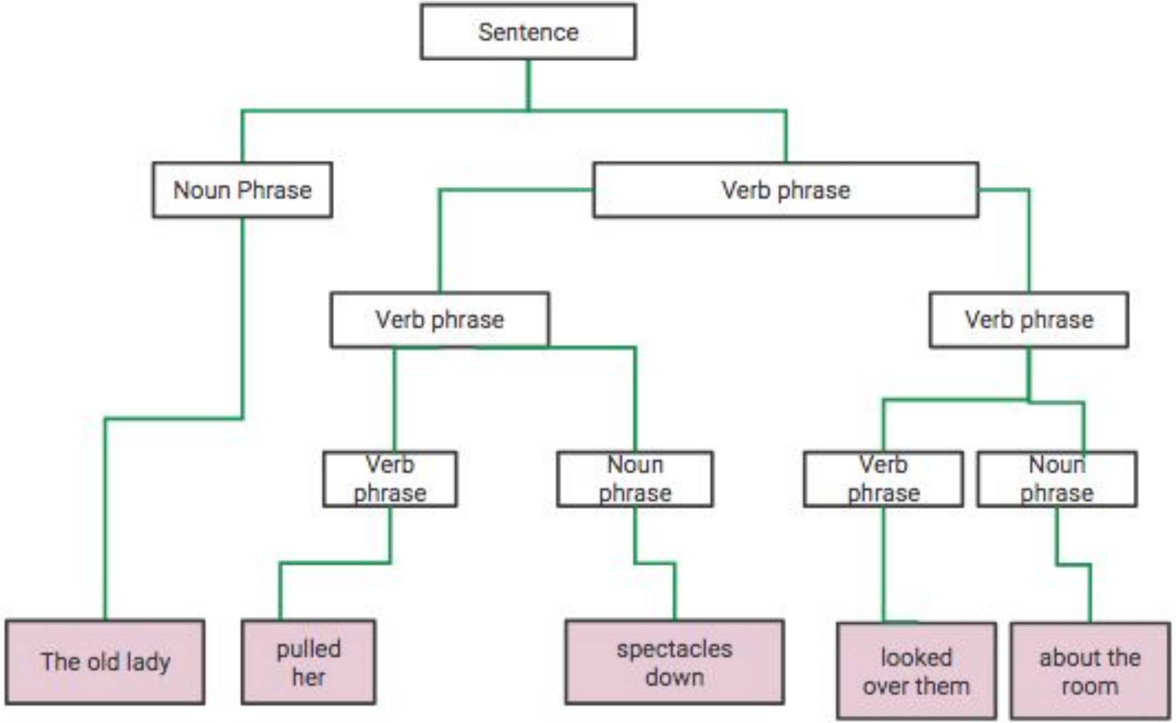
- ❑ **Rule Based or Logical methods** -classification and information retrieval
- ❑ **Probabilistic Models or Language Model** - QA,Information extraction
  - ❑ Documents are ranked based on the probability of the query *Q* in the document's language model.

Word Representation :

- ❑ **Can be Discrete - apple,boy,eat**
- ❑ **Distributional Approaches or Word2Vec---apple=[0.2,0.3,...]**
  - ❑ Assumes: Meaning is related to the context, Words that appear in same context are similar
  - ❑ Words are represented as continuous representation or embeddings of corpus vocabulary
  - – Requires large corpus to learn the relation between words

Applications:

TASK:	SIMPLE SOLUTIONS
a. Text Classification b. Information Retrieval c. Question Answering d. Information extraction e. Spelling correction f. Machine Translation	BOW or any n-gram using a classifier BOW with link ranking analysis BOW with if-else Information retrieval with rule based methods for association Character n-grams Rule based with POS tagging and semantic matching



Sentence

Noun Phrase

Verb phrase

Verb phrase

Verb phrase

Verb phrase

Noun phrase

Verb phrase

Noun phrase

The old lady

pulled her

spectacles down

looked over them

about the room

Simple Syntax Parse Tree showing Noun and Verb phrases in a sentence