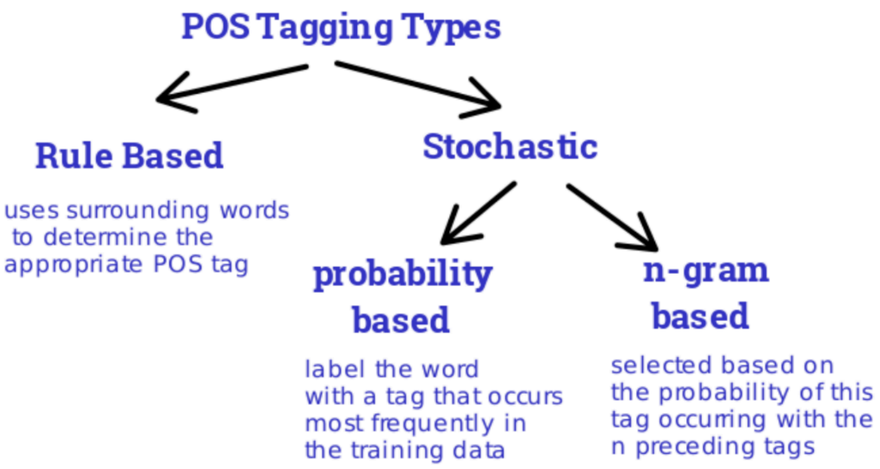
Lecture 5: POS Tagging and Topic Modelling

Part of Speech (POS) Tagging

process of classifying and labelling words into appropriate parts of speech, such as noun, verb, adjective, adverb, conjunction, pronoun and other categories.

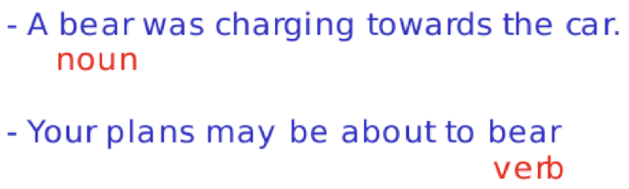

POS Tagging Types
Rule Based
uses surrounding words to determine the appropriate POS tag
Stochastic
probability based
label the word with a tag that occurs most frequently in the training data
n-gram based
selected based on the probability of this tag occurring with the n preceding tags

Examples of POS tags in NLTK (library of python):

- VBG – verb, present participle or gerund,
- PRP – pronoun, personal,
- NN – noun, common, singular or mass

Some Applications of POS Tagging

- Text-to-speech conversion
- Disambiguation of statements (check the word **bear**)



- A bear was charging towards the car.
noun

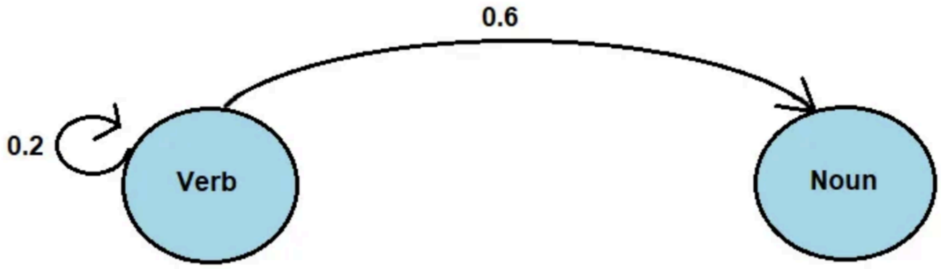
- Your plans may be about to bear
verb

- Named Entity Recognition, etc.

Markov Chains

Markov Model: representation of states and transitions to different states by assuming future states depend only on the current state

E.g.



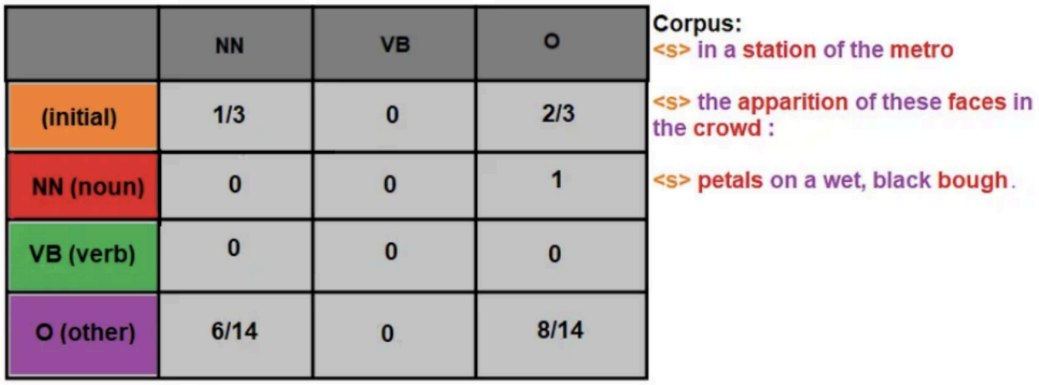
Current state → tail of the arrow

Future state → head of the arrow

the number on the arrow → Likelihood of tail followed by head

Transition Matrix

Another way to represent transition probabilities



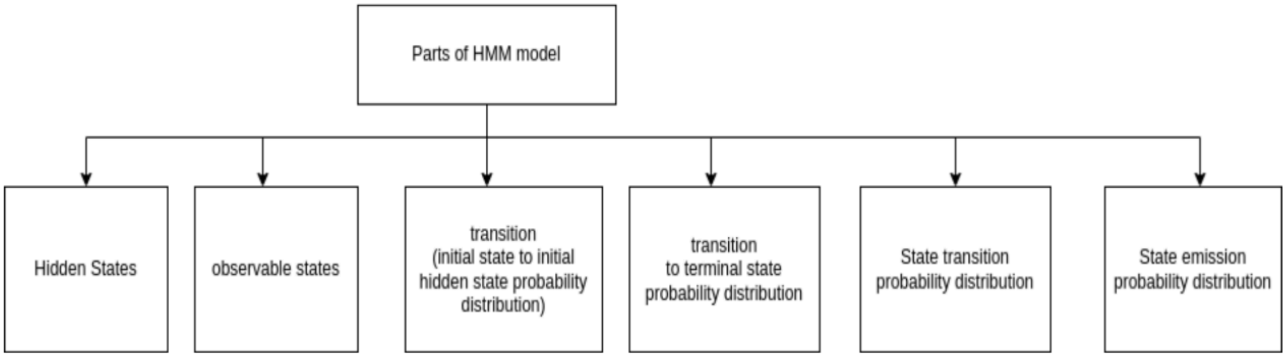
	NN	VB	O
(initial)	1/3	0	2/3
NN (noun)	0	0	1
VB (verb)	0	0	0
O (other)	6/14	0	8/14

Corpus:
<s> in a station of the metro

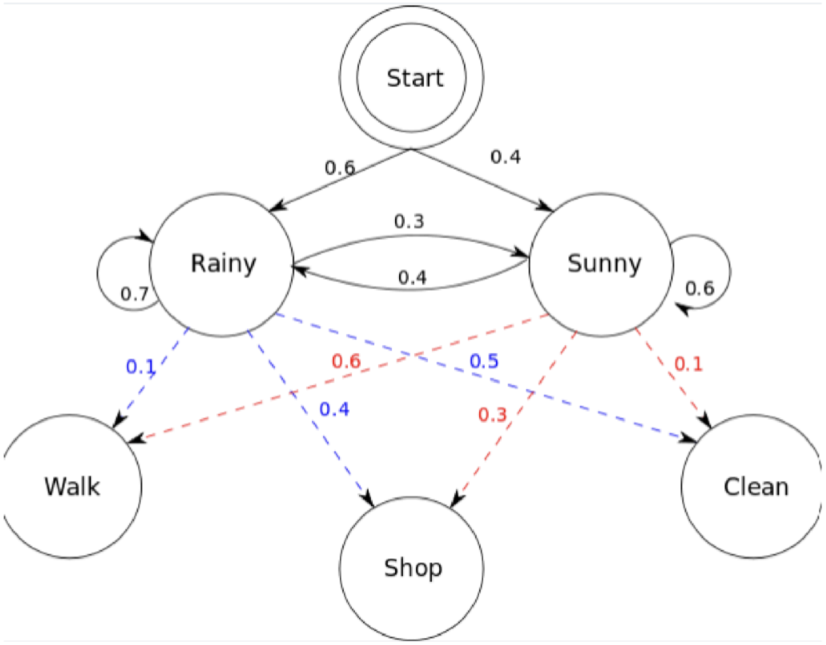
<s> the apparition of these faces in the crowd :

<s> petals on a wet, black bough.

Hidden Markov Model (HMM)



Example illustration



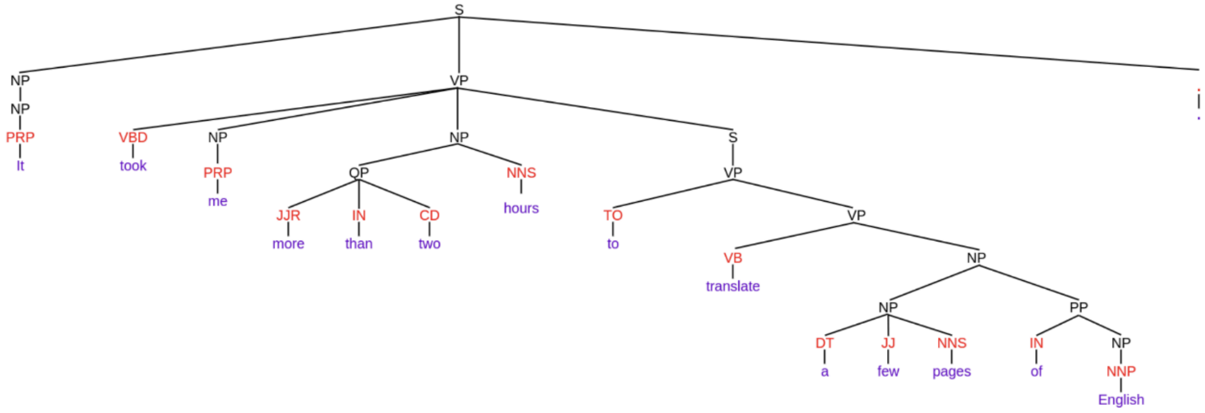
Dotted lines → transition to visible states

Observable (visible) states → words of the sentences

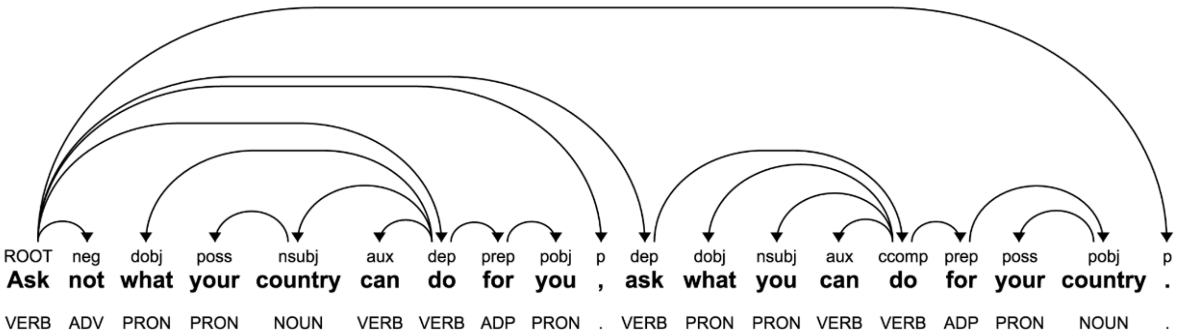
Hidden states → Parts of Speech

Constituency Parsing

Process of analysing the sentences by breaking down it into sub-phrases also known as constituents



Dependency Trees

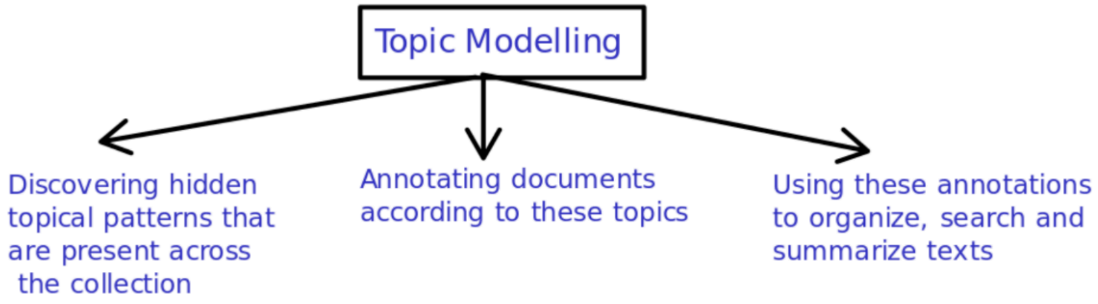


Arrows → parent-child relationship

Root → word with no arrow pointing towards it

Topic Modelling

Topics: Representative group of words in a large corpus.



Topic Modelling

Discovering hidden topical patterns that are present across the collection

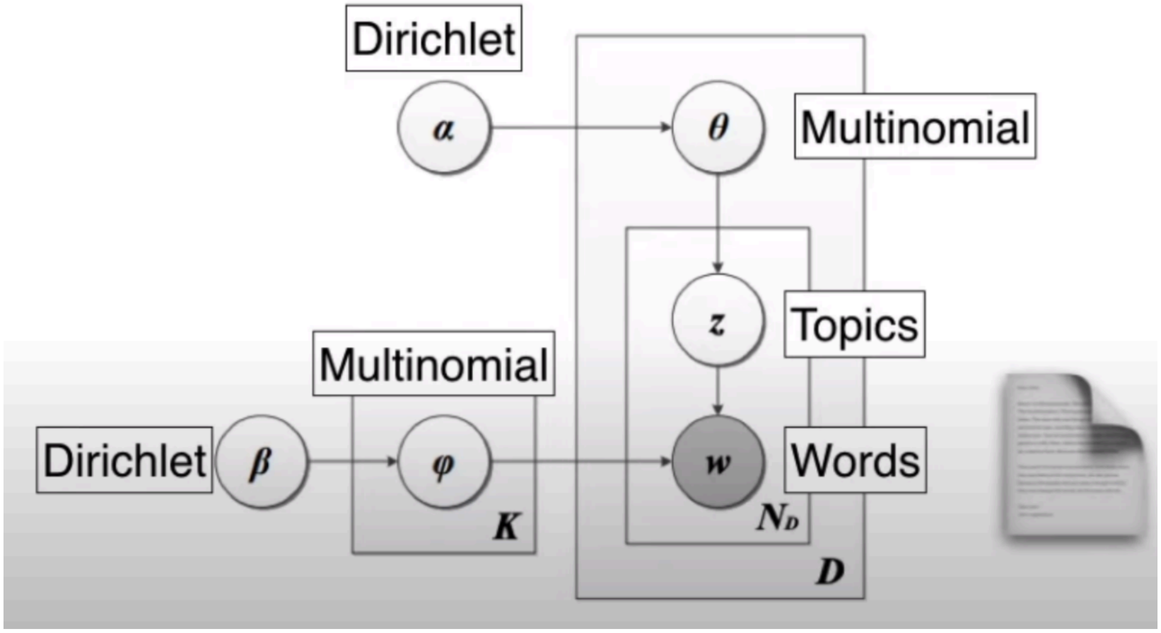
Annotating documents according to these topics

Using these annotations to organize, search and summarize texts

Latent Dirichlet Allocation (LDA)

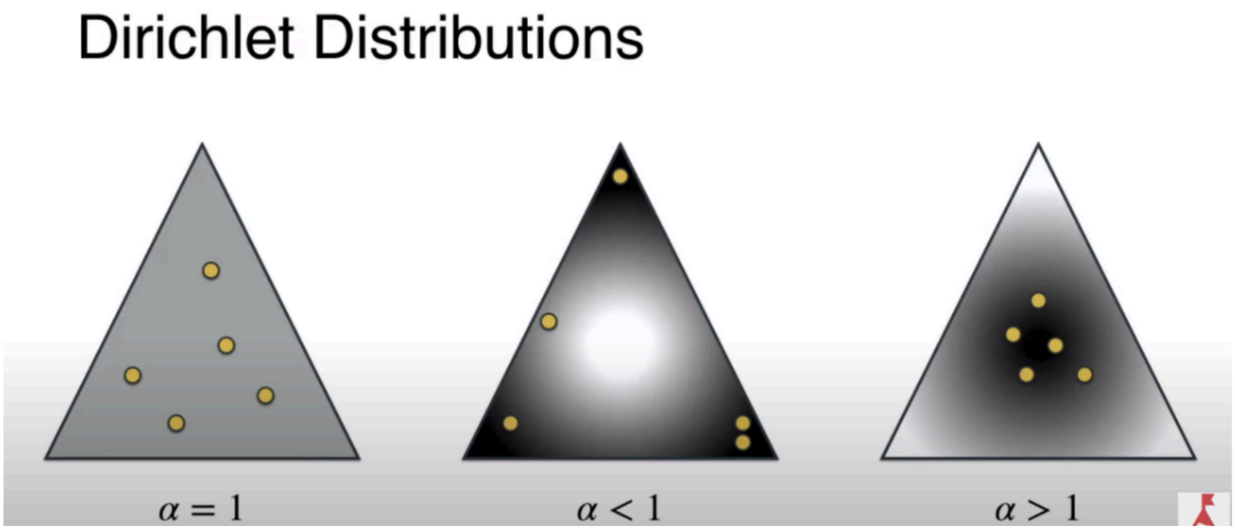
- LDA assumes that documents are composed of words that help determine the topics and maps documents to a list of topics by assigning each word in the document to different topics.
- While identifying the topics in the documents, it starts with random assignment of topics to each word and iteratively improves the assignment of topics to words through ***Gibbs sampling.***

Architecture of LDA:



LDA Hyperparameters

Hyper-parameter	usage
‘α’	document-topic density factor
‘β’	topic-word density factor
‘K’	number of topics to be considered (predefined)



Corners ⟶ Topics

Dots ⟶ Articles

The middle distribution represents the distribution of articles w.r.t. topics (as articles generally belong to a unique topic, lower probability of belonging to 2, and so on)

References:

Latent Dirichlet Allocation (Part 1 of 2)

Understanding Latent Dirichlet Allocation (LDA)