ALBERTA MACHINE INTELLIGENCE INSTITUTE

UNIVERSITY OF ALBERTA

NLP Workshop

HackED 2019

Luke Kumar
Machine Learning Scientist @ Amii

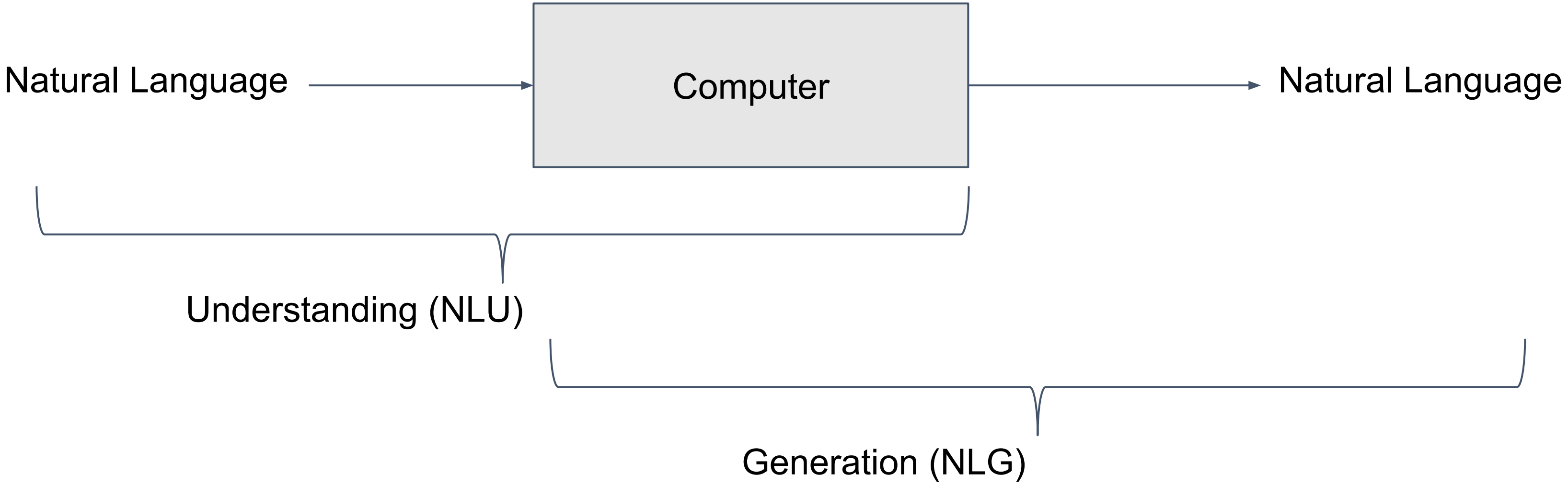
Agenda

1. NLP & History
2. Why NLP is hard?
3. Symbolic/Classical NLP
4. Statistical/ML NLP
    - Notebook Demos
5. Q&A

Note: Materials & Diagrams from Prof. Regina Barzilay's NLP [Course]() @ MIT (Advanced NLP)

Natural language processing

NLP - Building programs that can use NL as input and output



History of NLP

- First patents for translating machines were applied [mid-1930]
- Alan Turing published his famous article *Computing Machinery and Intelligence* which proposed what is now called the Turing test as a criterion of intelligence [1950]
- Noam Chomsky's Syntactic Structures universal grammar - a rule based system of syntactic structures [1957]
- Watson by IBM [2006]
- Word Embeddings [2013/2014]

Ref: https://en.wikipedia.org/wiki/History_of_natural_language_processing

Why NLP is hard?

- Ambiguity

  "Harry loves his mother and Hermione does too"

  - Harry and hermione love their own mothers
  - Hermione loves harry's mother
- Different types of ambiguities:
  - Acoustic (sound)
  - Syntactic (structure)
  - Semantic (meaning)
  - Discourse (multi-clause)
    - "The horse ran up the hill. **It** was very steep. **It** soon got tired"

How to solve ?

We need the:

1. Knowledge of the language
2. Knowledge of the world

Approaches:

1. Symbolic: Code all the rules into a program
2. Statistical: Learn language properties from examples

NLP in use ...

1. Language translation
2. Information extraction
    a. Search
3. Text summarization
4. Sentiment analysis
5. Text to Speech
    a. WaveNET
6. Chatbots - Conversational AI
    a. Alexa
    b. Google Home

…..

NLP in research - ACL 2019 tracks

- Dialogue and Interactive Systems
- Discourse and Pragmatics
- Document Analysis
- Generation
- Information Extraction and Text Mining
- Linguistic Theories, Cognitive Modeling and Psycholinguistics
- Machine Learning
- Machine Translation
- Multidisciplinary
- Word-level Semantics
- Multilinguality
- Phonology, Morphology and Word Segmentation
- Question Answering
- Resources and Evaluation
- Sentence-level semantics
- Sentiment Analysis and Argument Mining
- Social Media
- Summarization
- Tagging, Chunking, Syntax and Parsing
- Textual Inference and Other Areas of Semantics
- Vision, Robotics, Multimodal, Grounding and Speech

Classical/Symbolic NLP

Topics in symbolic NLP

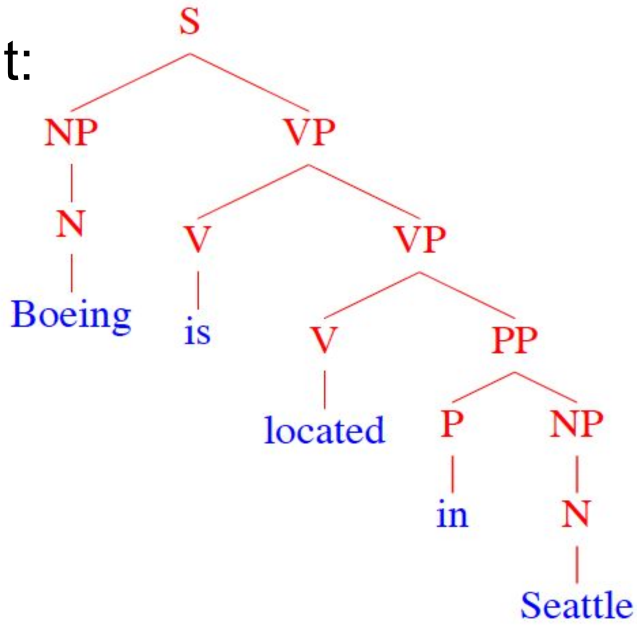
1. Parsing (we discuss)
2. Lexical semantics - Meanings of words
   a. WordNet - “mother”
3. Stemming and lemmatization
   a. am, are, is => be
   b. car, cars, car’s, cars’ => car
4. Named entity recognition (NER)
   a. Map text items to proper names (eg: people, location, organization)
5. ….

Parsing - syntactic structure

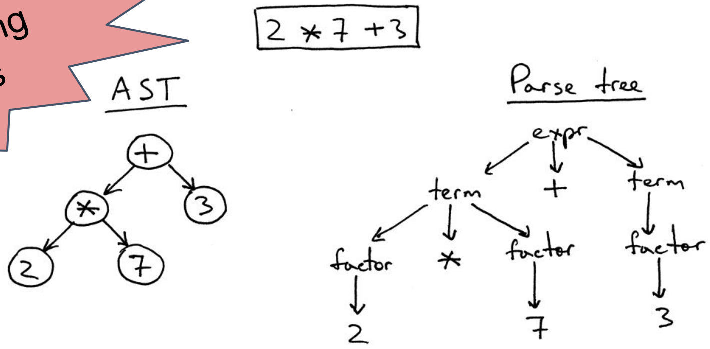
Input:

"Boeing is located in Seattle"

Output:



Programming Languages

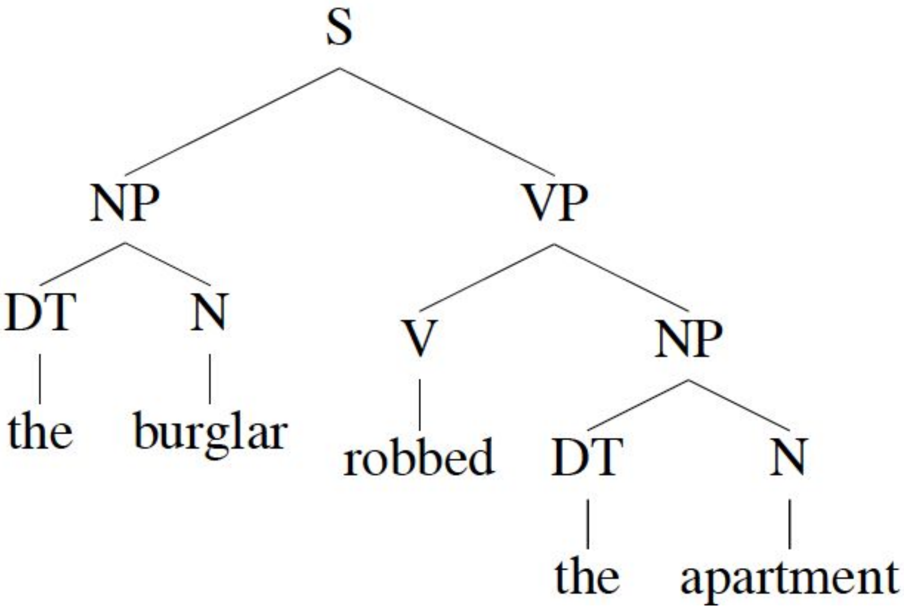


Ref: https://ruslanspivak.com/lsbasi-part7/

Parse tree

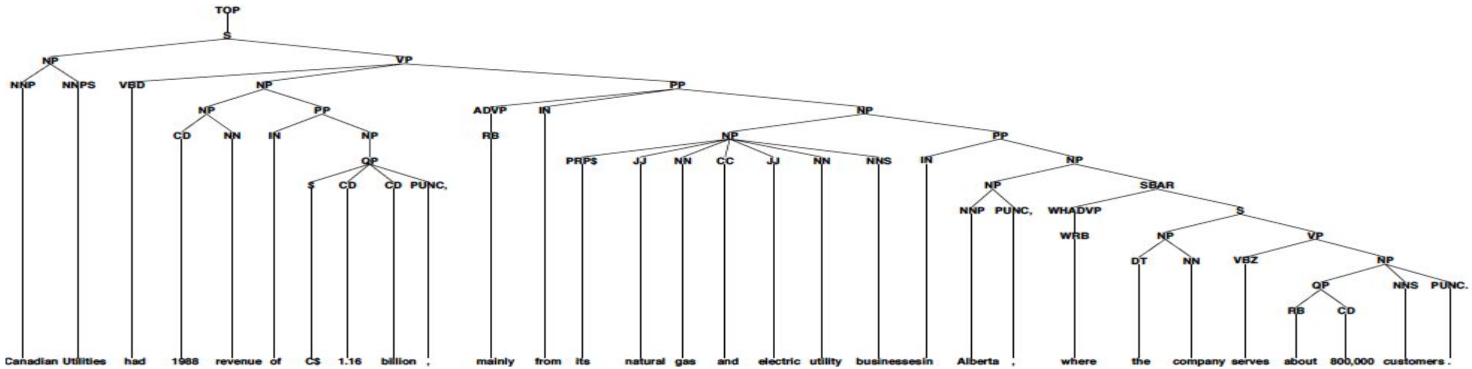
Parts of speech:

1. Words:
   a. N - noun
   b. V - verb
   c. DT - determiner
2. Phrases:
   a. NP - noun phrases
   b. VP - verb phrases
   c. S - Sentence



Penn Treebank

1. Major dataset for parsing experiments
2. ~ 50, 000 sentences along with trees



Canadian Utilities had 1988 revenue of C$ 1.16 billion , mainly from its natural gas and electric utility businesses in Alberta , where the company serves about 800,000 customers .

Grammars

1. Context-Free Grammars (CFG)
   a. Chomsky Normal Form (CNF)
2. Probabilistic CFG

$$N = \{\mathrm{S}, \mathrm{NP}, \mathrm{VP}, \mathrm{PP}, \mathrm{DT}, \mathrm{Vi}, \mathrm{Vt}, \mathrm{NN}, \mathrm{IN}\}$$
$$S = \mathrm{S}$$
$$\Sigma = \{\mathrm{sleeps}, \mathrm{saw}, \mathrm{man}, \mathrm{woman}, \mathrm{telescope}, \mathrm{the}, \mathrm{with}, \mathrm{in}\}$$

$R =$

S	$\Rightarrow$	NP	VP
VP	$\Rightarrow$	Vi	
VP	$\Rightarrow$	Vt	NP
VP	$\Rightarrow$	VP	PP
NP	$\Rightarrow$	DT	NN
NP	$\Rightarrow$	NP	PP
PP	$\Rightarrow$	IN	NP

Vi	$\Rightarrow$	sleeps
Vt	$\Rightarrow$	saw
NN	$\Rightarrow$	man
NN	$\Rightarrow$	woman
NN	$\Rightarrow$	telescope
DT	$\Rightarrow$	the
IN	$\Rightarrow$	with
IN	$\Rightarrow$	in

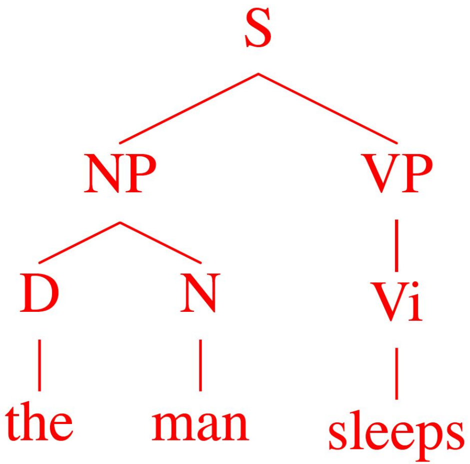
CFG - “the man sleeps”

$R =$

S	$\Rightarrow$	NP	VP
VP	$\Rightarrow$	Vi	
VP	$\Rightarrow$	Vt	NP
VP	$\Rightarrow$	VP	PP
NP	$\Rightarrow$	DT	NN
NP	$\Rightarrow$	NP	PP
PP	$\Rightarrow$	IN	NP

Vi	$\Rightarrow$	sleeps
Vt	$\Rightarrow$	saw
NN	$\Rightarrow$	man
NN	$\Rightarrow$	woman
NN	$\Rightarrow$	telescope
DT	$\Rightarrow$	the
IN	$\Rightarrow$	with
IN	$\Rightarrow$	in

Derivation	Rules
S	S -> NP VP
NP VP	NP -> DT NN
DT NN VP	DT -> the
the NN VP	NN -> man
the man VP	VP -> Vi
the man Vi	Vi -> sleeps
the man sleeps	



Statistical NLP

ML applications in NLP

Topics in statistical NLP (we discuss)

- Sentiment analysis
  - Positive vs negative polarity
- Language model
  - Probability distribution of a natural language
- Word embedding
  - Representing words as numerical vectors
- Topic model
  - Categorizing document collections

Other topics

- Statistical Machine Translation (SMT)
- Tagging (eg: POS tagging)
  - Hidden Markov Models (HMMs)
  - Conditional Random Forests (CRFs)
- Recurrent neural networks
  - Sequence to Sequence tasks (eg: Translation)
- Duplicate detection
- Other embeddings
  - Document2vector
  - Character2vector
- Conversational AI
- …..

Text to numeric representation

- Bag of words
  a. Absence and presence
  b. Frequency
  c. Term frequency inverse document frequency (TFiDF)
- N-grams
  a. Unigrams = Bag of words
- Embeddings
  a. Word embeddings
  b. Document embeddings
  c. Character embeddings

Sentiment analysis

- AKA Opinion mining
  - Polarity: positive or negative
- Sentiment Classification
  - Polarity Data 2.0 - Movie reviews
- Sentiment Lexicons
  - Subjectivity Lexicon
  - Bing Liu Opinion lexicon
  - SentiWordNet

Ref: https://web.stanford.edu/class/cs124/lec/sentiment.pdf

Demo

Sentiment_Analysis Notebook

Language model (Autoregressive models)

- Learn a probability distribution

$$\sum_{x \in \mathcal{V}^*} \hat{P}(x) = 1, \quad \hat{P}(x) \geq 0 \text{ for all } x \in \mathcal{V}^*$$

$$\hat{P}(\text{the}) = 10^{-12}$$
$$\hat{P}(\text{the fan}) = 10^{-8}$$
$$\hat{P}(\text{the fan saw Beckham}) = 2 \times 10^{-8}$$
$$\hat{P}(\text{the fan saw saw}) = 10^{-15}$$

Language model - n-grams

- Trigram (triplets) Model

  $P(w_i \mid w_{i-2}, w_{i-1})$

  eg:

  $$P(\text{"well"} \mid \text{"all"}, \text{"is"}) = \frac{Count(\text{all, is, well})}{\text{Count(all, is)}}$$

$$P(w_1, w_2, \ldots, w_{T-1}, w_T) = \prod_{t=1}^{T} P(w_t | w_{t-1}, w_{t-2}, \ldots, w_1)$$

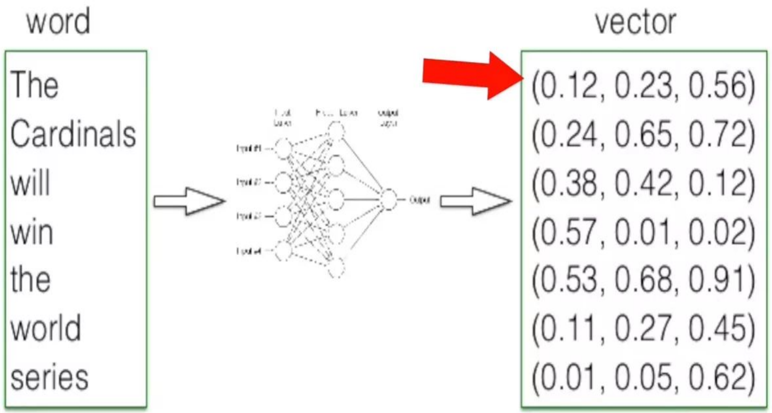
the	cat	sat	on	the	mat	$P(w_1)$
the	**cat**	sat	on	the	mat	$P(w_2 \| w_1)$
the	cat	**sat**	on	the	mat	$P(w_3 \| w_2, w_1)$
the	cat	sat	**on**	the	mat	$P(w_4 \| w_3, w_2, w_1)$
the	cat	sat	on	**the**	mat	$P(w_5 \| w_4, w_3, w_2, w_1)$
the	cat	sat	on	the	**mat**	$P(w_6 \| w_5, w_4, w_3, w_2, w_1)$

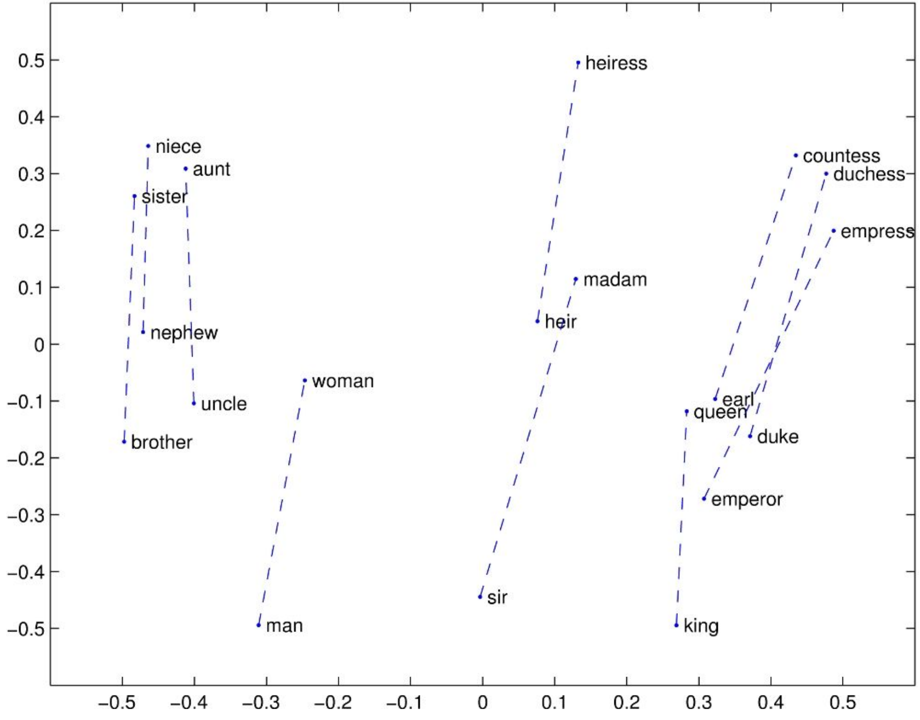
From Unsupervised Deep Learning tutorial @ NeurIPS 2018

Word embeddings

- Representations of NL
    a. Bag of words
    b. N-grams
- Embedding vectors - *Map words into vectors of real values*
    a. Embeddings are influenced by the context
    b. Embeddings try to capture the meaning using the context
- How embeddings are learnt:
    a. Language Model
        - Word2Vec (Google)
    b. Co-occurrence Matrix
        - GloVe (Stanford)

Word embeddings cont.





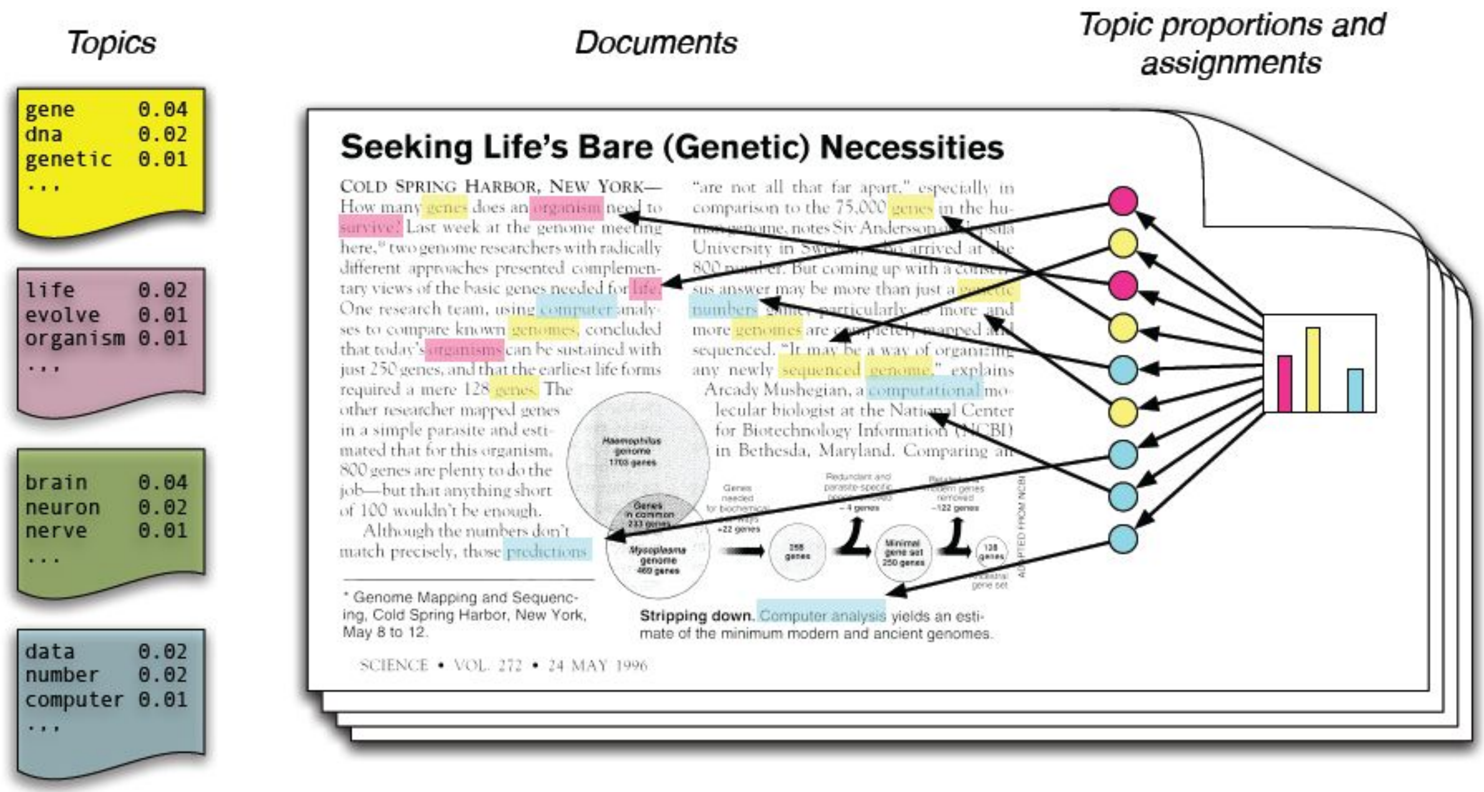
Demo

Word_Embeddings Notebook

Topic models

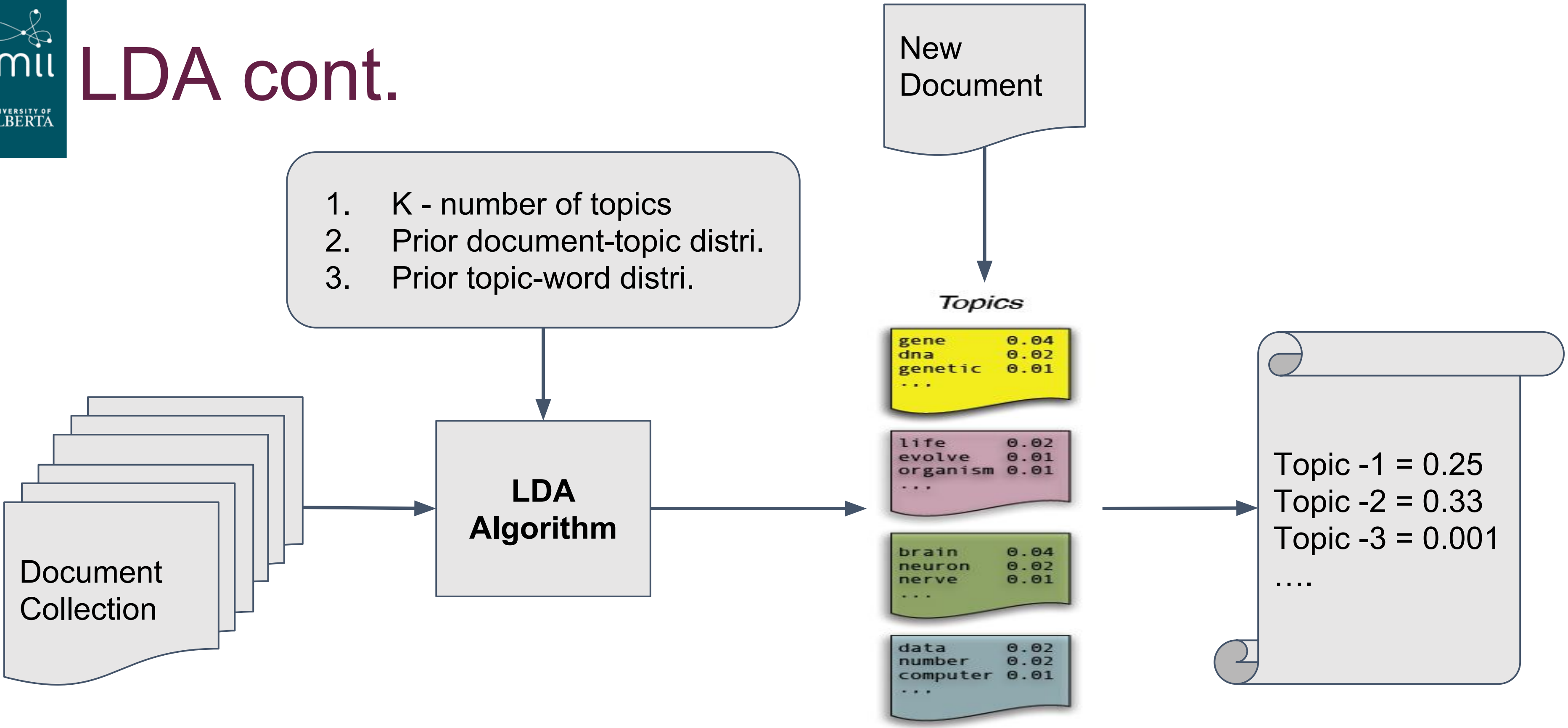
- A statistical model to learn intrinsic topics in a collection of documents
- Several models are proposed
  - LDA - Latent Dirichlet Allocation (Probabilistic)
  - LSA - Latent Semantic Analysis (SVD)
  - PLSA - Probabilistic Latent Semantic Analysis (Probabilistic)
- Helps to cluster a collection of documents
  - Soft clustering
  - Mostly interpretable (Probabilistic models)

LDA



Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Seeking Life's Bare (Genetic) Necessities
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

LDA cont.


New Document
1. K - number of topics
2. Prior document-topic distri.
3. Prior topic-word distri.
Topics
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Document Collection
LDA Algorithm
Topic -1 = 0.25
Topic -2 = 0.33
Topic -3 = 0.001
….

Demo

Topic_Models Notebook

Questions ?

amii.ca