Lecture 4

QM 701: Advanced Data Analytics

Fuqua School of Business

2024

Lecture Outline

- Topic Modeling (LDA and LSI)
- POS Tagging, NER, and Dependency Parsing

Topic Modeling

Two Paradigms of Machine Learning in NLP

- Supervised Learning
  - Training a model on a labeled dataset with the correct output
  - Example 1: Naïve-Bayes (Class 1)
  - Example 2: Feedforward Neural Networks (Class 2)
- Unsupervised Learning
  - Training a model on data without any labels
  - Example 1: word2vec (Class 3)
  - **Example 2: Topic Modeling (this class)**
  - Example 3: Language Models (Class 5 and 6)

Introduction to Topic Modeling

Topic modeling is a technique that automatically discovers abstract "topics" within a text corpus and classifies the documents based on these identified topics.

Business Applications:

Example: Topic Modeling for MyChart Messages

Topic	Top Keywords	Interpretation	Prevalence
1	Injection, treatment, hip, knee, weekend, pain, problem	Symptoms	7.4%
2	Fax, referral, bill, follow, pick, receive, correct, see, information	Administrative	15.3%
3	Bill, pick, tonight, supplement, vitamin, problem, thanks, soon, lol, week	Miscellaneous	15.3%
4	Generic, Raleigh, specialty, caremark, insurance, pharmacist, approve, drug, pay	Prescription	7.7%
5	Referral, colonoscopy, ultrasound, town, mammogram, procedure, book, surgery	Referral, procedure scheduling	7.1%
6	Testosterone, thyroid, study, biopsy, culture, result, fax, cholesterol, tsh, ultrasound	Test results	9.1%
7	Aspirin, drug, dosage, rx, level, pill, capsule, tablet, antibiotic, mcg, supplement, potassium	Medication question	7.7%
8	Asap, referral, fax, insurance, explain, record, name, procedure, company, discuss	Administrative	6.1%
9	Merry Christmas, xmas, holidays, January, season, wonderful, lveoly, happy, enjoy	Holidays	1.5%
10	Fax, referral, copy, vaccine, fmla, record, receive, report, pay, asap	Administrative	4.5%

From “COVID-19 and the Digital Health Transformation” by Wosik and Si, 2020

LDA and LSI Topic Models

LSI – Latent Semantic Indexing

- Simple and Fast
- Harder to interpret and use by humans
- Based on word counts and linear algebra

LDA - Latent Dirichlet Allocation

- Slower and more difficult to implement
- Easier to interpret and use by humans
- Based on statistical inference (iterative / algorithmic)

Model Selection for Topic Modeling

Model selection plays a critical role in the effectiveness of machine learning applications, especially in topic modeling. Key considerations include:

- Choice of topic modeling algorithms
  - e.g., choosing from LDA, LSA/LSI or others
- Determining the number of topics and tuning hyperparameters
- Preprocessing steps
  - Lemmatization
  - Common/uncommon words removal

Evaluating LDA Topic Models

- Perplexity score helps to quantifies how effectively a LDA model is representing the corpus.
  - However, we should not choose models by relying only on perplexity because:
    - Choosing a model based on only perplexity can result in models without meaningful insights
    - Low perplexity scores do not necessarily equate to more interpretable or useful models
- Alternative model evaluation approaches:
  - Inspecting the key words in each topic
  - Evaluate how well model categorize specific documents
  - Utilize alternative performance metrics such as coherence scores

POS Tagging, NER, and Dependency Parsing

Introduction to spaCy

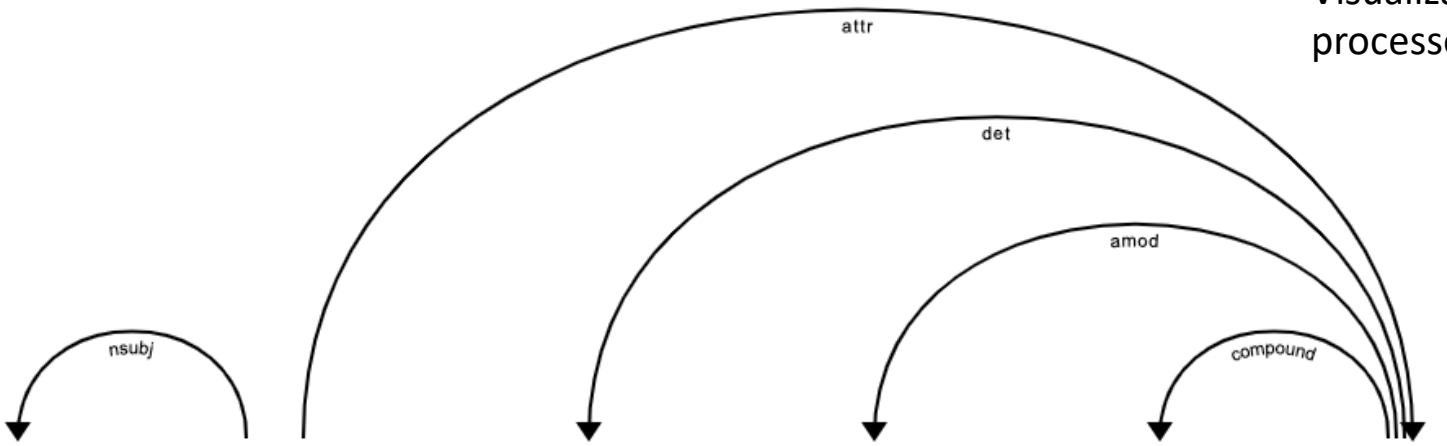
- An open-source library for advanced NLP tasks such as part-of-speech (POS) tagging, named entity recognition (NER), dependency parsing, etc
- Efficient and fast (especially compared to LLMs)
- Designed for practical, industrial-strength NLP tasks
  - In contrast, the nltk and genism libraries are more suitable for research
  - Supports many different languages and a wide range of NLP tasks

To learn more about Spacy, see https://spacy.io/usage/spacy-101

Part-of-Speech (POS) Tagging

- POS Tagging is the process of labeling each word in a sentence with its appropriate part of speech (e.g., noun, verb, adjective, etc.).
- Example sentence: 'Spacy is a useful software library.'

Visualization for the sentence processed by Spacy using Displacy



Spacy	is	a	useful	software	library.
NOUN	AUX	DET	ADJ	NOUN	NOUN
NN	VBZ	DT	JJ	NN	NN

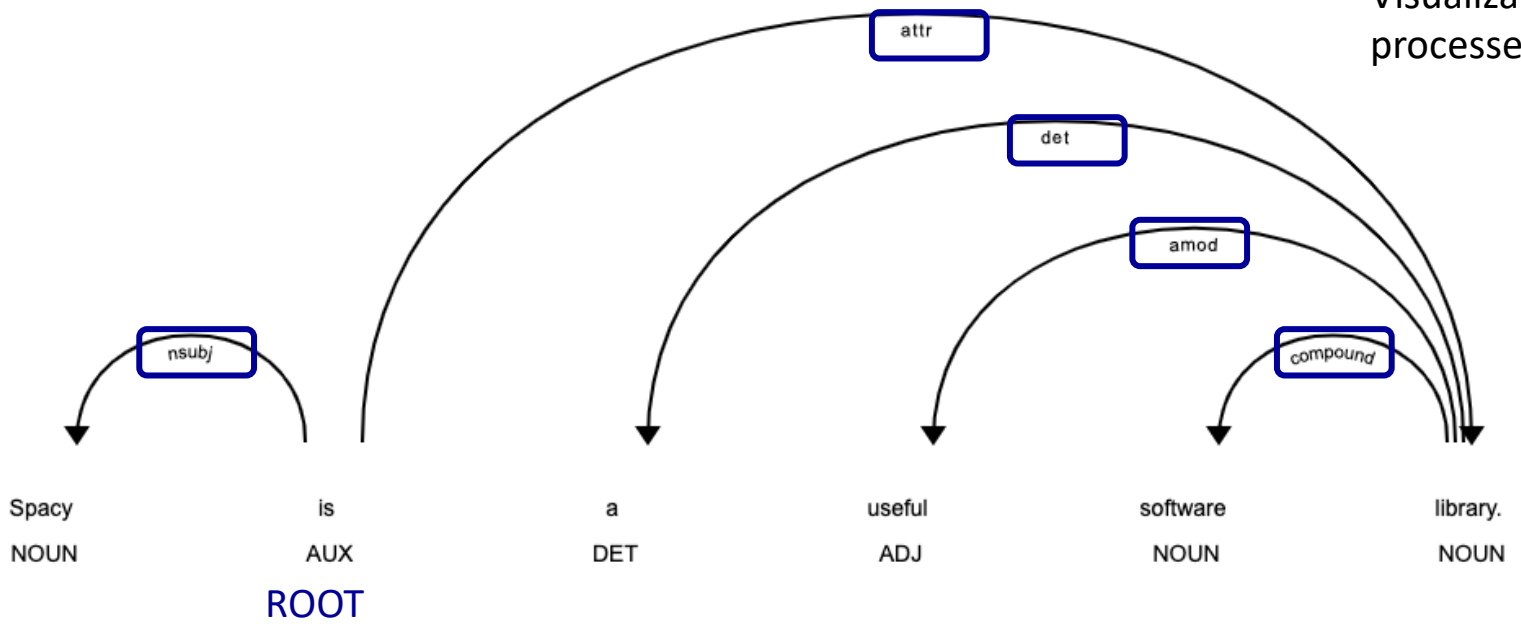
"coarse-grained" POS Tag

"fine-grained" POS Tag

Dependency Parsing

- Identify syntactic structure that shows how words relate to each other
- Represents “head” words and their syntactic “dependents”
- Basis for tasks that needs to understand the grammatical structure

Visualization for the sentence processed by Spacy



Dependency Relations

Named Entity Recognition (NER)

- NER classifies tokens in text into predefined entities.
- Identifies and extracts information about proper nouns, such as people, organizations, locations, etc.
- spaCy utilizes built-in, pre-trained statistical models to predict entities, some of the entity types in spaCy includes:
  - PER (people, including fictional)
  - GPE (countries, cities, states)
  - ORG (organizations, companies, agencies)
  - LOC (locations)

Detailed Explanation to Dependency Parsing Tags for Our Example

- “Spacy”: subject of the sentence. It is marked as ‘nsubj’ (nominal subject), implying this is the noun doing the action of the verb or the noun that the sentence is about
- “is”: the main verb of the sentence. In the dependency tree, it marked as ‘ROOT’, as the central node that other parts of the sentence are connected to.
- “a”: an article that modifies “library”. It is labeled as ‘det’ for determiner, which is a type of modifier that typically precedes nouns.
- “useful”: is an adjective that modifies “library”. It is tagged as ‘amod’, indicating that it is an adjectival modifier giving more information about the noun.
- “software”: modifies "library" to specify the kind of library being discussed. It is tagged as ‘compound’ since it forms a compound noun with "library."
- “library”: the main noun in the noun phrase “a useful software library”. It is tagged as ‘attr’ (attribute) which is dependent on ‘ROOT’ and serves to attribute a property to the subject, “Spacy”.

Homework 4

- Q1: Model Training. You will train two LDA models, one with 5 topics and another with 10 topics. (10 points)
- Q2: Model Evaluation. You will calculate the perplexity of each model and do the model selection. (15 points).
- Q3: Model Inspection. You will find the topic words of each model and compare these two models. (10 points)
- Q4: Model Test with Classic Books. You will test the model with several books and interpret the results (20 points)
- Q5: Conceptual Questions on Topic Modeling (20 points)
- Q6: Parsing with spaCy (15 points)
- Q7: Named Entity Recognition (NER) (10 points)