# CS 7337 – Natural Language Processing Final Exam

# By-Rashmi Patel

**Instructions:** Clarity of answers is more important than length of answers. Although not required (unless indicated otherwise), feel free to use graphs, charts, visuals, et al in your answers if you feel these artifacts can help support your answers. There are no bonus points for using these artifacts. **Submit your answers in PDF or Word document format.**

**Due date:** See course wall announcement.

**Q1.**

1. **[5 pts] What is Distributional Hypothesis in the context of distributional semantics?**

Give a short explanation with some examples.

According to Rubenstein & Goodenough (1965), the distributional hypothesis in the context of distributional semantics refers to the concept where words which are similar in meaning occur in similar contexts.

For example, if we want to find a word like ‘article’ then we can look for the words that occur in similar contexts, such as ‘journal’.

I was reading an article which provides information about COVID-19 detection in laboratory.

A journal is going to be published in IEEE regarding detecting COVID-19 using chest X-ray.

1. **[5 pts] Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) are two widely used techniques for topic modeling. Give a short overview of the two approaches and any similarities/differences between them.**

**Latent Semantic Analysis (LSA)-** LSA is one of the organic topic modelling techniques. The concept of LSA is to find the latent hidden terms which correlates semantically to form topics. Topics are unlikely to share the words because the words that are close in meaning will occur in similar pieces of context. The LSA learning make use of matrix decomposition on the document-term matrix using Single value decomposition. In other words, it begins with a large term-document matrix which usually populated with tf-idf values. After assigning words to topics, it then makes use of topic-to-topic matrix to determine what will maximize the variance between documents.

The steps to work with LSA are:

* Step 1: Create a term-document matrix (using term count or TF-IDF): rows are unique words, columns represent documents
* Step 2: Use Singular Value Decomposition (SVD) to reduce number of rows but preserve column similarity because the matrix can be very sparse and noisy o
* Step 3: Compare documents by cosine similarity between two vectors formed by two columns.

**Advantages of LSA:**

* Can cluster the words and documents in the space, so we can retrieve.
* Trying to solve the problems of polysemy and synonymy.
* Error reduction by dimension reduction

**Disadvantages of LSA:**

* Not giving well defined probabilities
* Lack of interpretable embeddings
* Requires large dataset of documents and vocabulary to get accurate results

Latent Dirichlet Allocation (LDA)- LDA is one of the organic topic modelling techniques. The concept of LDA is that each item of the collection (document) is modelled as a finite mixture over an underlying set of topic probabilities and each topic is then modelled as an infinite mixture over and underlying set of word probabilities. The assumption in LDA is that each document is generated by a statistical process. LDA is used in finding the weight of connections between documents and topics and between topics and words. The steps involved in LDA is shown below:

![Table

Description automatically generated with medium confidence]()

Advantages of LDA:

* Works well with large corpus
* Does not suffer overfitting issues
* Can be embedded in other complicated models
* Noise reduction is possible by dimension reduction

Disadvantages of LDA:

* Fixed k, the number of topics should be set in advance.
* Uncorrelated topics

**Comparison**: LDA is akin to LSA except in the way that it utilizes probability distributions over words, rather than a topic-topic matrix, to guide the assignment of words to topics.

![Diagram

Description automatically generated]()

# Q2.

1. **[5 pts] You are a Data Scientist for an e-commerce site for electronics which also**

supports 3rd party sellers. You would like to build a system to find and match the same products that sellers on your website sell so that you can present them in a single product page. You decide to use product titles to compute product similarity. Which similarity metric, Jaccard or Cosine, would you use and why?

For the given problem, it is better to use the Cosine Similarity. Cosine similarity is a better option because the scenario asks for finding and matching the same products using product titles. Cosine similarity of two documents is defined as the cosine angle formed between the two vectors. It can be represented as:

![Text

Description automatically generated]()

As we know that while building the system for matching the same products, it is possible that the product title might have duplicate names. Therefore, using cosine similarity for this problem will help because every duplicate does matter in determining similarity whereas Jaccard similarity is a case where duplication of words does not matter.

Therefore, we will use Cosine similarity for this problem.

1. **Consider the following table which lists electronic items for sale on two ecommerce shopping websites. Products in row -1 are the same product, row-2 are different TV models of the same brand and row-3 are different products.**

|  |  |
| --- | --- |
| **Product Title 1**  **(Site 1)** | **Product Title 2**  **(Site 2)** |
| 50 Inch Class H6570G 4K Ultra HD Android Smart TV with Alexa Compatibility 2.5” 2020 Model Black  Silver White HDR LED | Hisense H6570G |
| QN75Q90TAFXZA crystal 2.5” Quantum  LCD | Samsung crystal UN55TU8000FXZA QLED |
| EGLF2 50 Ultra Full Motion Articulating TV Wall Mount Bracket swivel full | VIZIO EGLF2 |

[10 pts] Considering your answer to 2a) will your similarity calculation approach work on this dataset? Explain with examples.

Row 2 with same company but different product (TV): After performing calculation for cosine similarity, it is found that 0.2242 similarity was there between the 2 different TV products for same company Samsung. Below image shows the detailed calculation:

Text, letter

Description automatically generated

Doing the same calculation for Row 1 with same product, 0.1957 cosine similarity was found and for Row 3 with completely different product, 0.204 cosine similarity was found.

Looking at the results it seems that cosine similarity approach works with this given problem.

[10 pts] Suppose that you are given IDF scores for all tokens (see Table below). Can this help you come up with a better approach for computing title similarity? Explain with examples.

|  |  |
| --- | --- |
| **Product Title 1**  **(Site 1)** | **Product Title 2**  **(Site 2)** |
| 50(6.3) Inch(8) Class(8.5) H6570G(10.2)  4K(9.4) Ultra(6.6) HD(5.7) Android(2.6)  Smart(6.1) TV(3.9) with(4) Alexa(6.9) Compatibility(15.6) 2.5”(5.7) 2020(6.8) Model(12.6) Black(6.8) Silver(7.8) White(12.6) HDR(12.2) LED (6.9) | Hisense(9.5) H6570G(10.2) |
| QN75Q90TAFXZA(13.7) crystal(11.3)  2.5”(5.7) Quantum(7.8) LCD(6.8) | Samsung(8) crystal(11.3) UN55TU8000FXZA(16.5) QLED(4) |
| EGLF2(15.6) 50(6.3) Ultra(6.6) Full(5.6)  Motion(6.7) Articulating(2.6) TV(3.7) Wall(8.5) Mount(9.5) Bracket(11) swivel (8.5) full (5.6) | VIZIO(10) EGLF2(15.6) |

Row 2 with same company but different product (TV): After performing calculation for cosine similarity, it is found that 0.27 similarity was there between the 2 different TV products for same company Samsung. Below image shows the detailed calculation:

Text

Description automatically generated with medium confidence

Doing the same calculation for Row 1 with same product, 0.21 cosine similarity was found and for Row 3 with completely different product, 0.48 cosine similarity was found which seems odd than what we saw previously.

Looking at the results it seems that cosine similarity approach works with this given problem.

# Q3.

a. [10 pts] Recommender systems are a subtype of information filtering systems that help users discover new and relevant items by presenting items like their previous interactions or preferences. Some famous examples of recommender systems are Amazon’s “Books you may like” and Netflix’s “Because you watched” carousels.

You are building a recommender system for your food delivery service startup and have data on co-purchases for food items f1, f2, . . ., fn (for example, food item f1 is commonly bought together with food item f4). How can you use techniques such as Word2Vec to recommend similar items to users who may have bought or show interest in any one of the items?

The Word2Vec in Recommender systems allow us to predict the words according to the context that are present nearby to each word in an input sentence. Using the similarity and dissimilarity concepts, it helps to generate word vectors.

Word vectors in Word2Vec works on creating the distributed representation of words and phrases and their compositionality. The word vector means the corresponding element is set to 1 and all other elements are set to 0.

For example- if the vocabulary is given as- King, Queen, Man, Woman, Child then, the vector for Queen is represented as: 0|1|0|0|0

Distributed representation of a word = each word is represented by a distribution of weights across all elements “meaning of a word”.

b. [10 pts] Word2Vec implements two different neural models: skip-gram and continuous bag of words (CBOW). Briefly explain the differences between the two models. Under which circumstances would you prefer the skip-gram model over CBOW?

Continuous Bag of Words (CBOW) is an efficient method for learning the high-quality distributed vector. It contains a sliding window of context words where one-hot encoding is done on each word. It consists of single hidden and output layer. The goal of CBOW is to maximize the conditional probability of observing the actual output words, regarding the weights.

Skip-gram is opposite of CBOW model. It contains a focus word which is the single input vector, and the target context words are the output layer. The goal of Skip-gram is to minimize the summed prediction error across al the context words in the output layer.

We can say that both the Word2Vec models have their own benefits and significance. Therefore, it totally depends on the given problem of which model to be used.

# Q4.

You are building a product classification system for an online electronics store. The

system should classify an incoming stream of millions of products to one of the 3000+ leaf level product types in the taxonomy such as laptops, smart TVs, wireless headphones, car speakers, among others. The system should be very precise because it’s important to assign products to the right category to facilitate the customer shopping experience. Each instance in yourdataset has product title, description and image fields. See example below:



1. **[5 pts] What features would you use for your machine learning-based classifier?**
2. **[5 pts] Assume that you only have access to product titles in your dataset (i.e., you have less data to play with) instead of product titles, description and images. How will this affect feature engineering and the NLP pipeline for your classifier?**
3. **[10 pts] Obtaining training data is paramount for a large-scale classification system. You have a limited budget and can’t hire an army of analysts to manually label every single instance. Discuss some strategies for obtaining training data for the classifier.**
4. **[5 pts] How would you handle products that are misclassified?**

# Q5.

1. **[10 pts] Sentiment analysis: consider the following review of a restaurant:**

***“I took my father out for dinner to Le Bistro on New Year’s Eve. The décor and service were fantastic. We enjoyed the food, especially their French countryside specials and their Chardonnay collections. However, my father thought the menu prices were a bit on the high side. Valet parking was also expensive. Overall, we definitely recommend Le Bistro for special occasions!”***

***Overall rating: 8 stars out of 10***

***“***

Identify the opinion object(s), feature(s), opinion(s), opinion holder(s) and opinion time in this review.

1. **[10 pts] Design a sentiment analysis system for restaurant reviews (see example in 5a). Your answer should make use of the techniques discussed in class. The output of the system should assign a sentiment label of Positive or Negative to reviews.**