CS 7337 – Natural Language Processing
Final Exam

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.

a. [5 pts] What is Distributional Hypothesis in the context of distributional semantics? Give a short explanation with some examples.

b. [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.

Q2.

a. [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?

b. 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.**

[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)

Q3.

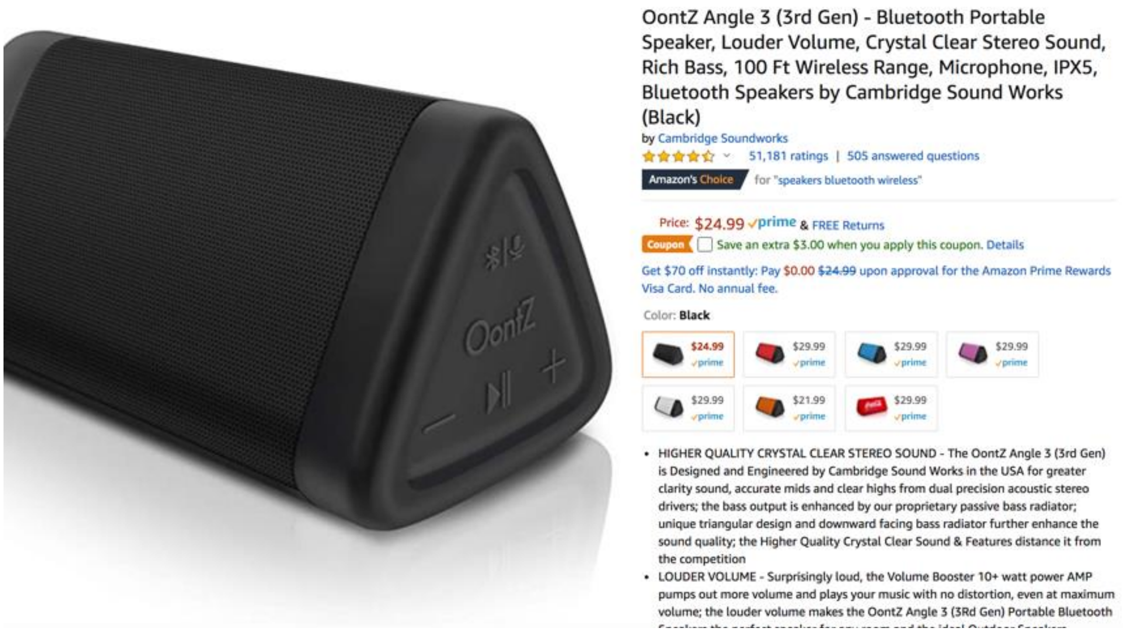
a. [10 pts] Recommender systems are a subtype of information filtering systems that help users discover new and relevant items by presenting items similar to their previous interactions or preferences. Some famous examples of recommender systems are Amazon's "Books you may like" and Netflix's "Because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

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?

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 your dataset has product title, description and image fields. See example below:



OontZ Angle 3 (3rd Gen) - Bluetooth Portable Speaker, Louder Volume, Crystal Clear Stereo Sound, Rich Bass, 100 Ft Wireless Range, Microphone, IPX5, Bluetooth Speakers by Cambridge Sound Works (Black)

by Cambridge Soundworks

★★★★☆ 51,181 ratings | 505 answered questions

Amazon's Choice for "speakers bluetooth wireless"

Price: $24.99 ✓prime & FREE Returns

Coupon □ Save an extra $3.00 when you apply this coupon. Details

Get $70 off instantly: Pay $0.00 ~~$24.99~~ upon approval for the Amazon Prime Rewards Visa Card. No annual fee.

Color: Black

$24.99 ✓prime	$29.99 ✓prime	$29.99 ✓prime	$29.99 ✓prime
$29.99 ✓prime	$21.99 ✓prime	$29.99 ✓prime	

- HIGHER QUALITY CRYSTAL CLEAR STEREO SOUND - The OontZ Angle 3 (3rd Gen) is Designed and Engineered by Cambridge Sound Works in the USA for greater clarity sound, accurate mids and clear highs from dual precision acoustic stereo drivers; the bass output is enhanced by our proprietary passive bass radiator; unique triangular design and downward facing bass radiator further enhance the sound quality; the Higher Quality Crystal Clear Sound & Features distance it from the competition
- LOUDER VOLUME - Surprisingly loud, the Volume Booster 10+ watt power AMP pumps out more volume and plays your music with no distortion, even at maximum volume; the louder volume makes the OontZ Angle 3 (3Rd Gen) Portable Bluetooth

a. [5 pts] What features would you use for your machine learning-based classifier?

b. [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?

c. [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.

d. [5 pts] How would you handle products that are misclassified?

Q5.

a. [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.

b. [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.