• This is a closed book and closed notes exam. No electronic aids or cheat sheets are allowed.

• There are 8 pages, 4 parts of questions (the last part is for bonus questions), and 115 total points in this exam.

• The questions are printed on both sides!

• Please carefully read the instructions and questions before you answer them.

• Please pay special attention on your handwriting; if the answers are not recognizable by the instructor, the grading might be inaccurate (NO argument about this after the grading is done).

• Try to keep your answers as concise as possible; grading is NOT by keyword matching.

| Total     | /115 |
Academic Integrity Agreement

I, the undersigned, have neither witnessed nor received any external help while taking this exam. I understand that doing so (and not reporting) is a violation of the University’s academic integrity policies, and may result in academic sanctions.

Signature: ________________________

Your exam will not be graded unless the above agreement is signed.
1 True/False Questions (12×3 pts)

Please choose either True or False for each of the following statements. For the statement you believe it is False, please give your brief explanation of it (you do not need to explain when you believe it is True). Two point for each question. Note: the credit can only be granted if your explanation for the false statement is correct.

1. Stemming reduces the vocabulary size.
   **True**

2. When we represent a text corpus with document-word adjacency matrix, removing tail words will reduce more storage space than head words.
   **True**

3. Words with high DF are more discriminative than those with low DF.
   **False, and Explain:** low DF words are usually more discriminative as they only appear in a handful of documents.

4. Vector space model is equivalent to the Bag-of-Word model.
   **False, and Explain:** Bag-of-Word model is just a special case of vector space model, where we treat individual words/N-grams as the bases.

5. A harmonic mean is used in the F1 score instead of arithmetic mean because of computational complexity.
   **False, and Explain:** We use harmonic mean to capture outlier in precision or recall.

6. Cosine similarity is preferred in vector space models because it is normalized with respect to document length.
   **True**

7. Interleaved test is more sensitive than A/B test because it is not biased towards any particular ranking algorithm.
   **False, and Explain:** Interleaved test is more sensitive because it provides all alternatives for users to compare.

8. If we could have a corpus with an infinite number of documents, smoothing is not needed when estimating those document language models.
   **False, and Explain:** Because of Zipf’s law, in single documents we still will not observe all words in the vocabulary, so that smoothing is still needed to estimate the document language model.

9. The goal of retrieval models we have learnt is to improve some specific IR evaluation metrics, such as NDCG and MAP.
   **False, and Explain:** The goal of any retrieval model is to help user fulfill their information need.

10. We usually make independence assumption for the purpose of reducing computational cost in retrieval models.
    **True**
11. \( \text{P@2=0.2} \) means in average we will have 20\% of chance to observe a relevant document at position 2 over all the queries from this particular ranking algorithm. 

\textit{False, and Explain: It should be at the first two positions.}

12. Given a very large IR evaluation collection, where System A achieves a MAP of 0.33 and System B achieves a MAP of 0.79, we can safely conclude that System B is significantly better than System A. 

\textit{False, and Explain: Statistical test is needed to confirm the statistical significance, no matter how large the average difference is (Or if you argue search quality should be also measured by other metrics, such as MRR or NDCG).}

## 2 Short Answer Questions (49 pts)

Most of the following questions can be answered by one or two sentences. Please make your answer concise and to the point.

1. Give one reason you might choose Bag-of-Words Representation over Inverted Index. Give one reason you would choose Inverted Index over the Bag-of-Words Representation. (6pts)

   - If we need to frequently scan through the words in particular documents.
   - If we need to frequently scan through the documents containing a particular word.

2. One of the data structures we originally discussed for indexing was a map from documents to linked lists of the words that occurred in the documents. Explain why Zipf’s law guarantees that the inverted index data structure will have better time complexity than this document-keyed map. (5pts)

   Due to Zipf’s law, a very large portion of words only appear in a handful of documents. As a result, in query time, most of documents will not contain the target query term, which leads to a high missing rate if we scan through the documents to match against that word.

3. List two major reasons that focused crawling is usually preferred than DFS or BFS based crawling strategies. (4pts)

   - It is impossible to exhaust the whole web during crawling
   - Not all documents are equally important

4. List three key elements in a classical IR evaluation framework. (6pts)

   - A document collection
   - A set of information needs expressed as queries
   - A set of relevance judgments on query document pairs
5. List three key elements shared by most of retrieval models (such as BM25 and language models) we have learned in class. (6pts)

- Term frequency
- Document frequency
- Document length normalization

6. Write down one of the techniques used in spelling correction, and give a short example. (4pts)

- Use edit distance to identify spelling candidates, such as “apple” to “applw”
- Consider the layout of a keyboard to identify spelling candidates, such as “apple” to “applw”
- Consider query context, such as “iphone case” to “iphone cease”
- Consider phonetic similarity between words, such as “herman” to “Hermann”

7. Write down the formula of NDCG and explain which parts of the formula correspond with which letters in the name. Then, explain why we have to normalize it, i.e. why is the DCG not good enough? (8pts)

\[ DCG = \sum_{i=1}^{n} \frac{2^{rel_i} - 1}{\log_2(1 + i)} \]

iDCG is the DCG calculated under the ideal ranking of documents for the same query, and NDCG is defined as the ratio between DCG and iDCG. N stands for the ratio-based normalization, D stands for the position-based discount (i.e., \( \log_2(1 + i) \)), C stands for the cumulation of gain obtained from all the returned documents (i.e., the summation), and G stands for the relevance-based gain (i.e., \( 2^{rel_i} - 1 \)). We perform normalization of DCG to reduce potential bias from queries associated with many relevant documents.

8. List three major assumptions of classical IR evaluation. (6pts)

- Top to bottom scanning of results
- Independent relevance judgements over documents
- All the users share the same relevance judgements

9. Give two reasons why Dirichlet Prior smoothing is preferred over Add-1 smoothing. (4pts)

- Not all the words are equally important
- Incorporate document length

\[ \text{DCG} = \sum_{i=1}^{n} \frac{2^{rel_i} - 1}{\log_2(1 + i)} \]
3 Essay Questions (15 pts)

The following question focuses on system/algorithm design. Please think about all the methods and concepts we have discussed in class (including those from the students’ paper presentations) and try to give your best designs in terms of feasibility, comprehensiveness and novelty. When necessary, you can draw diagrams or write pseudo codes to illustrate your idea. Ten points for each question.

- Twitter is building its next generation of tweet search engine. Engineers are arguing that classical IR techniques are totally sufficient to satisfy this goal so that there is no need for them to work overtime. As Twitter’s chief research scientist in IR, do you agree with them? If not, what are the major technical challenges? What has to be innovated and what has to be adapted in such a system? (*Hint: think about the system architecture of a classical IR system and the nature of Twitter.*)

   Important aspects in your answer should include: 1) dynamic indexing, given the emerging new tweets are generated every second; 2) document length normalization might not be essential, given the fixed maximum length of a single tweet; 3) freshness should be emphasized in result ranking in twitter; 4) collaborative ranking, social connections, social endorsement (e.g., retweet, favorite) needed to considered in ranking; 5) the language on twitter is changing, e.g., jargon, abbreviations and hashtags, and it increases the vocabulary gap.

4 Bonus Questions (15 pts)

All these questions are supposed to be open research questions. Your answers have to be very specific to convince the instructor that you deserve the bonus (generally mention some broad concepts will not count).

1. How could you predict a user’s next query? (10pts)

   This is an open research question, and I am expecting some cool new ideas that are beyond what we have discussed in class, e.g., using search logs, location, or friendship. Anything reasonable deserves the bonus points.

2. List your favorite (labeled as +) and disliked (labeled as -) aspects of this class. (5pts)

   Any constructive comments and suggestions deserve the bonus points.