CMU 67-300: Search Engines **Name:** _____
April 2017 **Andrew ID:** _____

This exercise contains 4 pages (including this cover page) and 6 questions.

1. **Document Representation I** We studied a series of text pre-processing decisions that need to be taken by the system architect. For each one of the following methods: (1) define it; (2) specify whether it is likely to *increase* or *decrease* the size of the indexed collection and justify you answer; (3) specify whether it is likely to *increase* or *decrease* the recall of your system and justify your answer.

   Lower casing all tokens

   Stemming

   Removing stop words

2. **Document Representation II** Typically, words have many different meanings. WordNet, for instance, describes 7 different meanings for *dog*. Suppose you created an optimal ninja method that always finds the correct meaning of a word. You want to use this perfect method to increase the precision of your system. Describe which modification you need to do when indexing (processing documents) and retrieving (processing a query). Why this optimal ninja method can increase the precision of your system?

3. **Retrieval Models** The main components of any retrieval method studied are **TF**, **IDF** and **document length normalization**. Explain each one of these components and how they are represented in **one** of the retrieval model seen (either VSM, LM or BM25).

4. **Relevance Feedback** Define what is Pseudo-Relevance Feedback. Using this technique, is it more likely to *increase* or *decrease* the **recall** of your system? Why?

   .....

   .....

   .....

   .....

   .....

5. **Language Models** Consider we have a collection with only the following 4 documents:
   Doc1: "So I am on BBC2 now telling Terry Wogan how I made it"
   Doc2: "What I made is unclear now but his deference is and his laughter is"
   Doc3: "My words and smile are so easy now"
   Doc4: "Yes It is easy now"

   Given the query "I made it" how the query-likelihood model would sort these documents? Do not apply use stemming nor smoothing. Remember: you should calculate $P(\text{“}Imadeit'' \mid Doc1)$, $P(\text{“}Imadeit'' \mid Doc2)$...and order them by score.

6. **Evaluation Metrics** There are many evaluation metrics for search engines (examples: **P@3**, **P@10**, **NDCG**, **MRR**, **MAP**). Please pick the most appropriate evaluation metric for each of the following search tasks. Justify your answer.

   A student searching for the answers of this exam on his mobile phone after this exam.

   A football fan searching for information and history for Fifa World Cup. Some of the returned pages provide a lot of relevant details, for example, team rankings, match scores, the latest news, etc. Some pages are just marginally relevant. Others are less interesting or irrelevant.

   A business man searching for the Al Jazeera homepage to see the news after his diner.