DATA20021 Information Retrieval
Exam, 9 March 2022, 16:00 to 18:30, Moodle
Lecturers: Dorota Glowacka and Simon Puglisi

There are a total of 40 points in this exam.

1. [*3+6+1 points*] Consider the following document collection:
   doc 1: The smallest planet is Mercury.
   doc 2: Jupiter, the largest planet!
   doc 3: Neptune is smaller than Uranus, but is heavier.
   doc 4: Uranus is smaller than Saturn.
   doc 5: Neptune is heavier than Uranus.
   doc 6: Jupiter > Saturn > Uranus > Neptune

   (a) Draw the term-document incidence matrix for the collection after normalizing by applying case folding and removing punctuation.

   (b) Draw the inverted index for the collection of documents after normalizing by applying case folding and removing punctuation. Include document frequencies and term frequencies in your picture.

   (c) What are the returned results for these queries:

      i. peace **AND** drug

      ii. new **AND NOT** (drug **OR** hope)

2. [*3 + 3 points*] Encode the sequence of integers 9, 29, 229 using

   (a) Elias γ codes;

   (b) Vbyte codes.

3. [*2 + 2 points*] Consider the table below of term frquencies for three documents from a collection of 906799 documents.

Table 1: Term frquencies and document frequencies.

term	Doc 1	Doc 2	Doc 3	df_t
slow	5	17	19	11816
this	30	33	67	606881
bird	0	6	9	1969
down	15	33	24	12523

   (a) First compute the idf values for each term and then compute the tf.idf weights for the terms for each document using the document frequency values (df_t column) in the table.

   (b) Compute the Euclidean normalized document vectors for each of the three documents above, where each vector has four components, one for each of the four terms.

4. [*5 + 5 + 5 points*] An engineering company with 10,000 employees worldwide has recently developed an in-house prototype search system to replace an old Unix system. The search system is used mostly by engineers who need to find precise information in technical manuals quickly. The engineers also use the system to keep up to date with the latest

developments in their specific fields and so occasionally browse through industry specific magazines and journals that the company has online access to. The information systems department of the company has asked you to evaluate the prototype search systems to ensure that it supports the needs of the engineers using it. One of the information systems specialists working for the company also asks you a number of questions about the evaluation of the prototype.

(a) They say that they have read in one of the information systems journals that commercial search engines often use A/B testing. They want to know if you are planning to use this method as well when evaluating their prototype system. Explain to them what A/B testing is and why it is appropriate/inappropriate for the problem at hand.

(b) They also say that they came across evaluation measures such as MAP and NDCG. They want to know how these two measures differ from more traditional measures of precision and recall and whether there is any advantage of using MAP and NDCG over precision and recall. They want to know if in your evaluation, you will be using any of the discussed methods (i.e. precision, recall, MAP, NDCG) and why/why not.

(c) The information system specialist thinks it would be a good idea to test the new system versus the old system before deploying it throughout the whole company. The information systems specialist suggests you should conduct a user study using 5 or 6 of the company secretaries working in head office of the company and the results analysed using statistical testing. If the results are significant, then you can move on to deployment. Is the system specialist right? Explain you answer.

5. [*5 points*] You are a product manager in an e-commerce company. The online shopping process of your website includes: entering the homepage, browsing products, adding products to the cart, initiating the purchase process, and completing a purchase. One day, a UI designer comes to you and suggests making the colour of the "buy now" button in the checkout page more prominent to attract customers' attention.

   You want to conduct an A/B test to understand if the change benefits your business. You selected revenue-per-user as the OEC and you want a 5% increase in revenue (practical significance level). Based on past experience, you assume that 5% of users who visit during the experimental period will end up purchasing, and those purchasing will spend 75EUR on average. The standard deviation is assumed to be 30EUR.

   As 5% of users spent 75EUR, the average revenue-per-user is 3.75EUR. Recall that $n = \frac{16\sigma^2}{\Delta^2}$ (under the condition two sample t-test, statistical power 80%, and statistical significance threshold 0.05). The number of users required for the A/B is therefore: $n = \frac{16\times 30^2}{(3.75\times 0.05)^2} = 409,000$.

   You have only two weeks to run the test. However, you notice that the estimated sample size is almost double the number of active users of the website for two weeks. Please specify what kind of strategies you could use to reduce the sample size.