## Information Organization Midterm Report (November 22, 2023)

Notice

- Write your name and student number.
- You can write answers either in English or Japanese.

**Question 1.**

| **term** | apricot | apple | applemango | appendix | apply | and |
|---|---|---|---|---|---|---|
| **document frequency** | 194 | 511 | 10 | 86 | 2540 | 50970 |
| **rank** | 2500 | 1000 | 50000 | 6000 | 200 | 10 |

Suppose that there are N=52000 documents in a document collection, and a dictionary is constructed on this collection. The above table shows six terms and their document frequencies in the dictionary. Here, the rank $r$ of term $t$ indicates that $t$'s frequency is $r$-th highest in the dictionary.

(1) How a very frequent term like "and" in the above is called?

(2) Compute IDF score of each term in the table.

(3) Suppose that an inverted index having non-positional posting lists is constructed using this dictionary, where each posting consists of document ID and its term frequency, and the postings are sorted by the descending order of term frequencies. Show an efficient method to return the top-100 documents ranked by TF-IDF score, for the Boolean query below, where '|' means OR, '&' means AND, and '^' means NOT. List up the names of the methods you used.

(^applemango | apple)&apply & apricot

Answer within 200 words.

(4) Estimate a function $F(r)$ that approximately computes term frequency $F(r)$ from its rank $r$.

(5) How the distribution of (3) is called?

(6) Show a space-efficient format of the dictionary, utilizing front coding, recoding their frequencies and pointers to their posting lists.

**Question 2.**

Suppose there is a collection of 10000 documents, including documents $d_1$, $d_2$ and $d_3$. Also, terms "**fresh**", "**fish**" and "**sushi**" appear in the collection and $d_1$, $d_2$ and $d_3$ with the frequencies shown below:

| term | collection frequency | document frequency | frequency in $d_1$ | frequency in $d_2$ | frequency in $d_3$ |
|---|---|---|---|---|---|
| **fresh** | 4000 | 2000 | 5 | 10 | 50 |
| **fish** | 2000 | 500 | 10 | 20 | 10 |
| **sushi** | 1000 | 200 | 10 | 0 | 20 |

(1) We want to rank the terms in the documents $d_1$, $d_2$ and $d_3$ by importance, using tf-idf score. Calculate the tf-idf score of each term, using ltn.lnc, namely logarithmic term frequency (1+$\log_{10}$ ($\mathrm{tf}_{t,d}$)), idf ($\log_{10}$ N/$\mathrm{df}_t$) document frequency. Here, use $\log_{10} 2 = 0.30$ and $\log_{10} 5 = 0.70$.

(2) We want to measure similarities among $d_1$, $d_2$ and $d_3$, by cosine similarity.

(2-1) Show the term frequency vectors of $d_1$, $d_2$ and $d_3$ on the three terms, where the term frequency is logarithmic (1+$\log_{10}$ ($\mathrm{tf}_{t,d}$)), and length-normalized.

(2-2) Compute the cosine values for each pair of the three vectors, and determine which document pair is most similar.

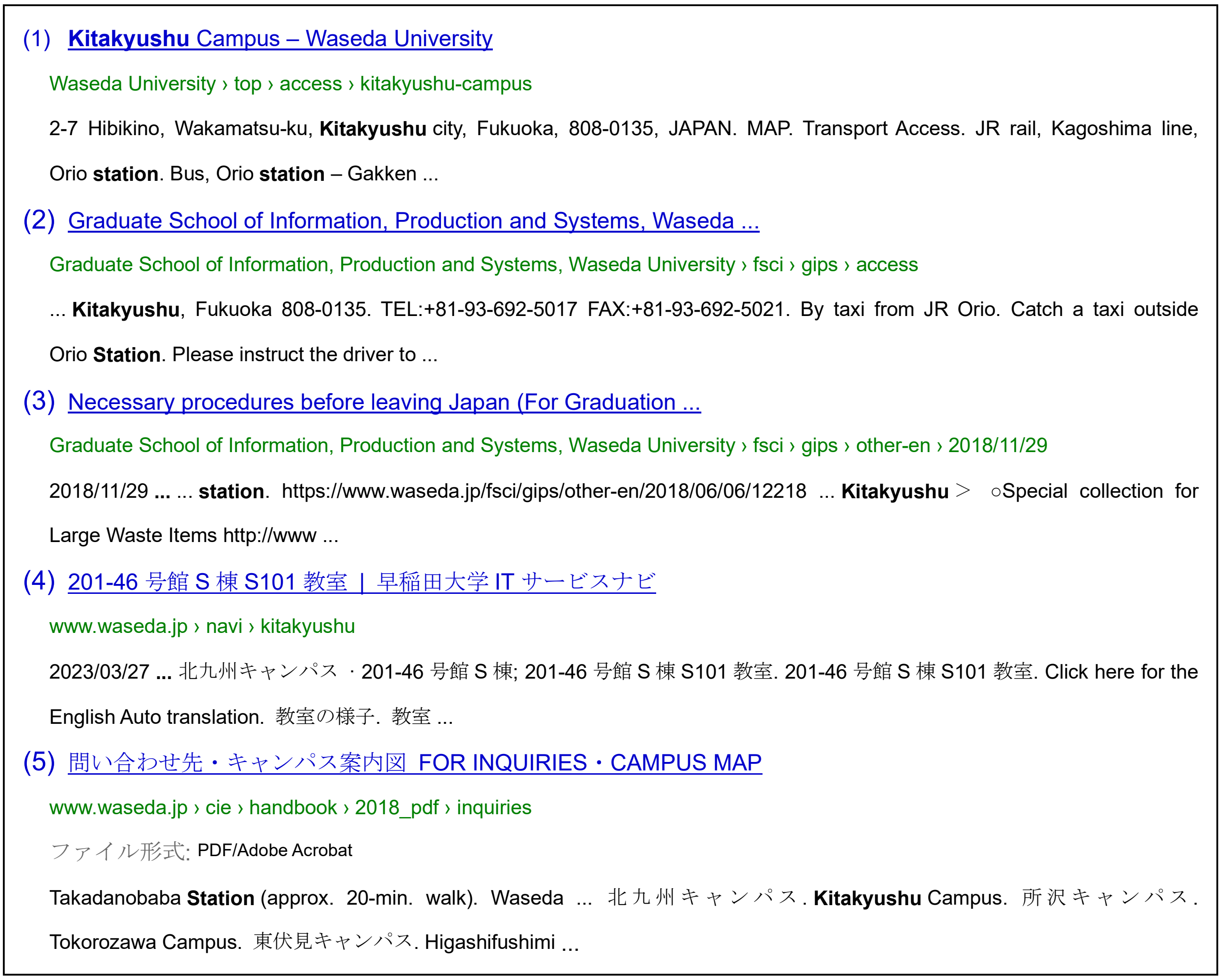
Question 3.

| document | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Alice | 4 | 0 | 2 | 1 | 3 | 1 | 2 | 5 | 3 | 2 | 1 | 2 |
| Bob | 3 | 2 | 5 | 3 | 4 | 0 | 3 | 3 | 0 | 2 | 3 | 1 |
| SysA | 3 | 1 | 6 | 1 | 0 | 1 | 3 | 6 | 2 | 4 | 1 | 0 |
| SysB | 5 | 1 | 4 | 3 | 3 | 1 | 5 | 6 | 0 | 3 | 0 | 3 |

Alice and Bob ranked documents 1-12 on how each document is relevant to query “seasonal greetings” and scored each document from 5 (highest) to 0 (lowest). Let $s(X,i)$ denote the score user or system $X$ on document $i$. For example, $s(\text{Alice},1) = 4$.

1. Define that Alice and Bob *agree* on document $i$ if $s(\text{Alice},i) \geq 3$ and $s(\text{Bob},i) \geq 3$ or $s(\text{Alice},i) < 3$ and $s(\text{Bob},i) < 3$. Otherwise, they have no agreement on $i$. Now, how you evaluate the level of agreement between Alice and Bob?
2. Suppose that information retrieval systems SysA and SysB are tested on query “seasonal greetings” and they returned ranking on documents 1-12, as shown in the same table. Also, define the *average score* of document $i$ as $avg(i) = [s(\text{Alice},i) + s(\text{Bob},i)]/2$. Now suppose that document $i$ is relevant to the query if $avg(i) \geq 3$. Using this relevance, compute the precision, recall, and F1-score of SysA and SysB.
3. Now determine which system in SysA or SysB is producing a ranking that is closer to the ranking by the average score $avg(i)$ of the 12 documents. Here, consider using precision@K, and recall@K, with K varying from 1 to 12. Here, precision (resp. recall@K) with respect to a ranking is the precision (resp. recall) when documents ranked higher or equal to K are regarded as relevant.
4. Discuss measurement of similarities of rankings other than the method used in previous question (within 100 words).

Question 4.



(1) **Kitakyushu** Campus – Waseda University

Waseda University › top › access › kitakyushu-campus

2-7 Hibikino, Wakamatsu-ku, **Kitakyushu** city, Fukuoka, 808-0135, JAPAN. MAP. Transport Access. JR rail, Kagoshima line, Orio **station**. Bus, Orio **station** – Gakken ...

(2) Graduate School of Information, Production and Systems, Waseda ...

Graduate School of Information, Production and Systems, Waseda University › fsci › gips › access

... **Kitakyushu**, Fukuoka 808-0135. TEL:+81-93-692-5017 FAX:+81-93-692-5021. By taxi from JR Orio. Catch a taxi outside Orio **Station**. Please instruct the driver to ...

(3) Necessary procedures before leaving Japan (For Graduation ...

Graduate School of Information, Production and Systems, Waseda University › fsci › gips › other-en › 2018/11/29

2018/11/29 **...** ... **station**. https://www.waseda.jp/fsci/gips/other-en/2018/06/06/12218 ... **Kitakyushu** ＞ ○Special collection for Large Waste Items http://www ...

(4) 201-46 号館 S 棟 S101 教室 ｜ 早稲田大学 IT サービスナビ

www.waseda.jp › navi › kitakyushu

2023/03/27 **...** 北九州キャンパス ・201-46 号館 S 棟; 201-46 号館 S 棟 S101 教室. 201-46 号館 S 棟 S101 教室. Click here for the English Auto translation. 教室の様子. 教室 ...

(5) 問い合わせ先・キャンパス案内図 FOR INQUIRIES・CAMPUS MAP

www.waseda.jp › cie › handbook › 2018_pdf › inquiries

ファイル形式: PDF/Adobe Acrobat

Takadanobaba **Station** (approx. 20-min. walk). Waseda ... 北九州キャンパス. **Kitakyushu** Campus. 所沢キャンパス. Tokorozawa Campus. 東伏見キャンパス. Higashifushimi ...

The above is the top-5 result of the search query by terms “kitakyushu” and “station”.
https://www.waseda.jp/top/en/search?q=kitakyushu+station

(1) Explain where the following notions are used in this top-5 result.

(a) snippet. (b) metadata. (c) dynamic summary.

(2) Argue what ranking strategy could have been used to generate this top-5 result (within 200 words).

Notice: question (1) is related to result summaries in Chapter 8, p.50, which is not covered in the lecture.