The Information Retrieval Exam is a normal paper exam in a classroom at NWU.

You cannot use calculators, mobile phones or books in the exam. You must bring your NWU ID card to the exam.

Students are not allowed to leave the exam during the last half-hour. This is to avoid disrupting other students who are finishing their work. You may leave early, up until the last half-hour.

Please do \*not\* fold your exam paper or answer sheets. The answer sheets will be scanned after the exam so they need to be unfolded to go through the machine easily.

The exam consists of five questions, each of 20 marks.

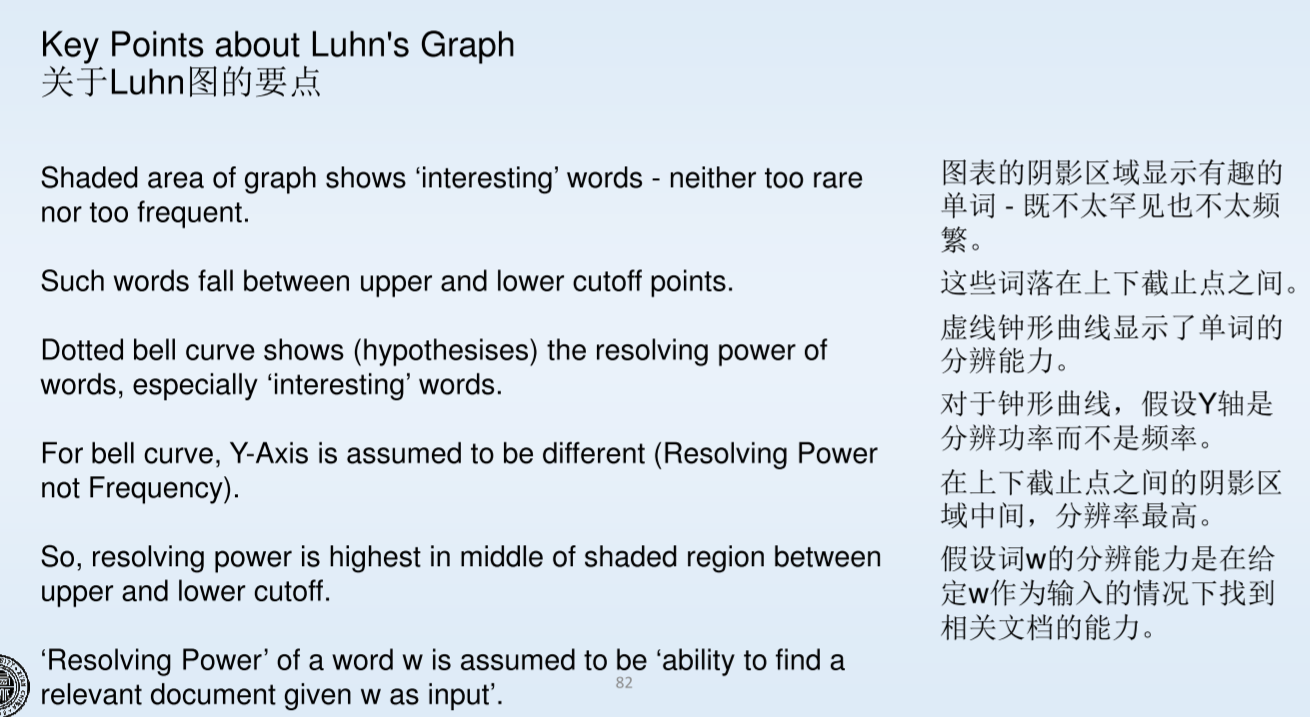
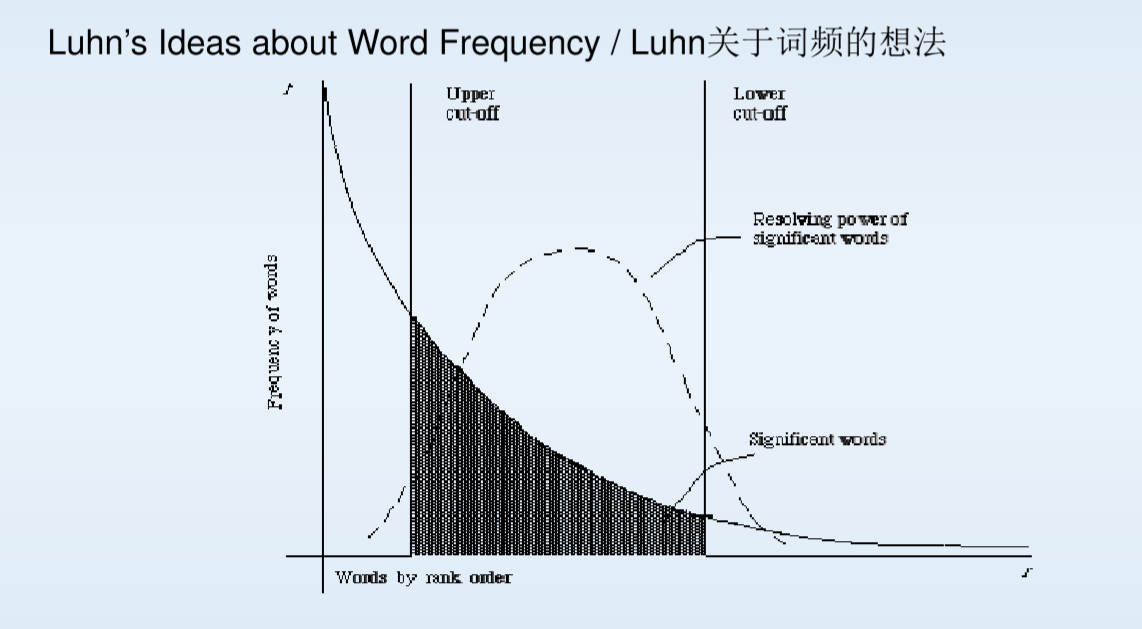
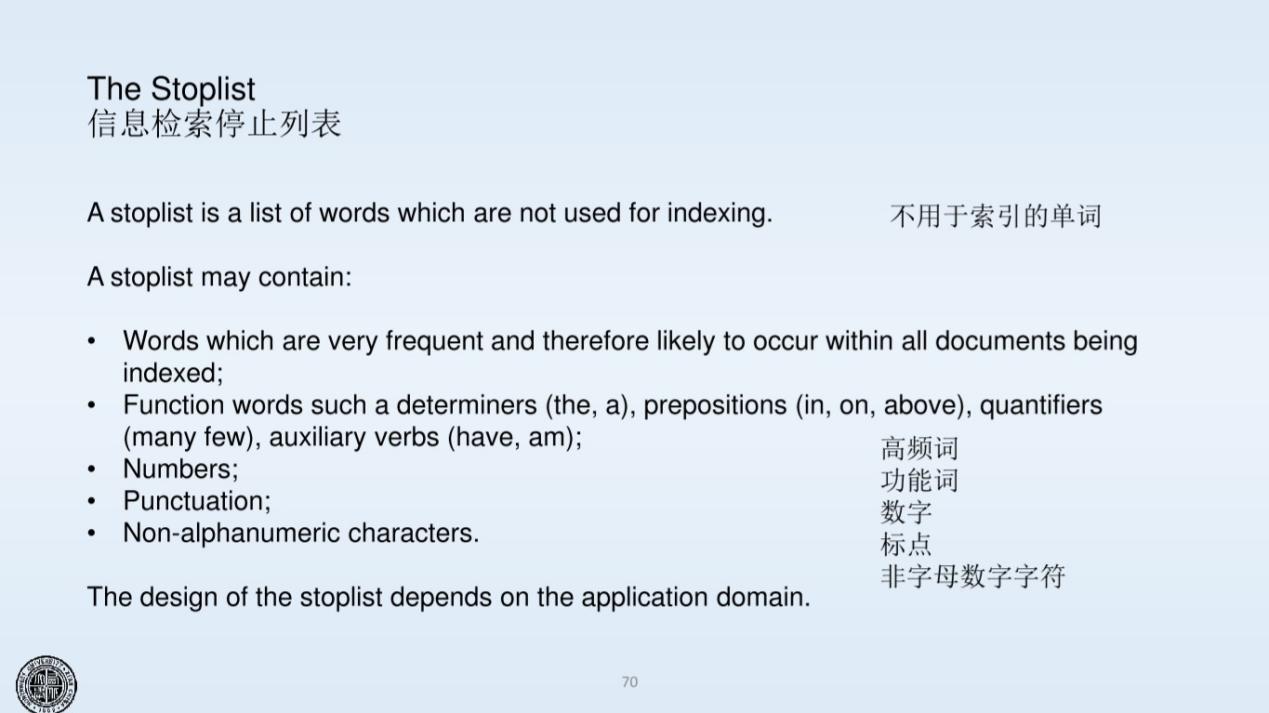
Please read through the exam paper in the first ten minutes. If you have any questions about the English, you may ask the invigilators. They will contact me as necessary for clarification.

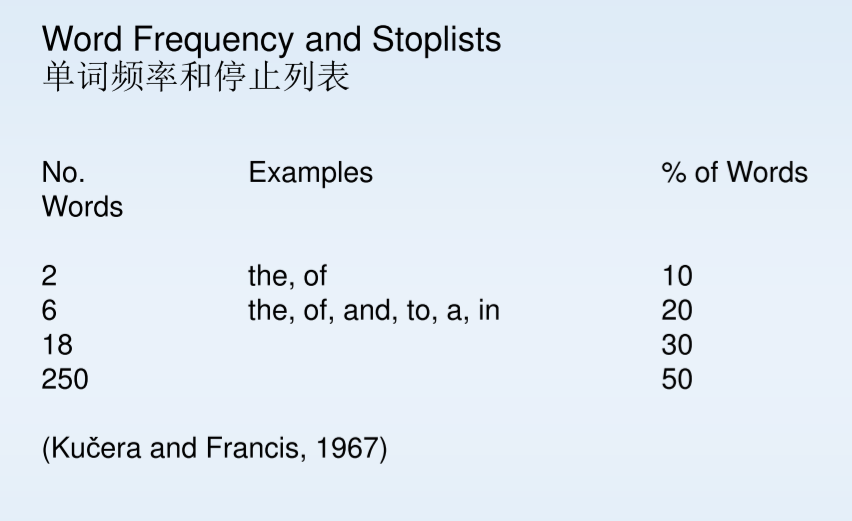
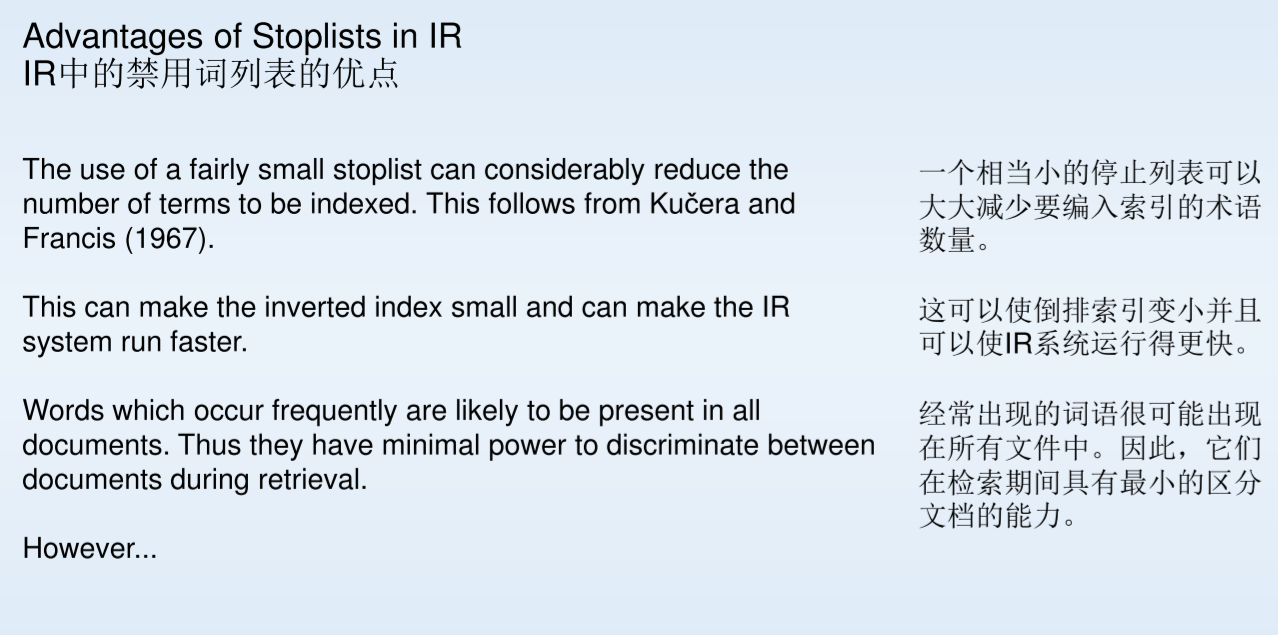
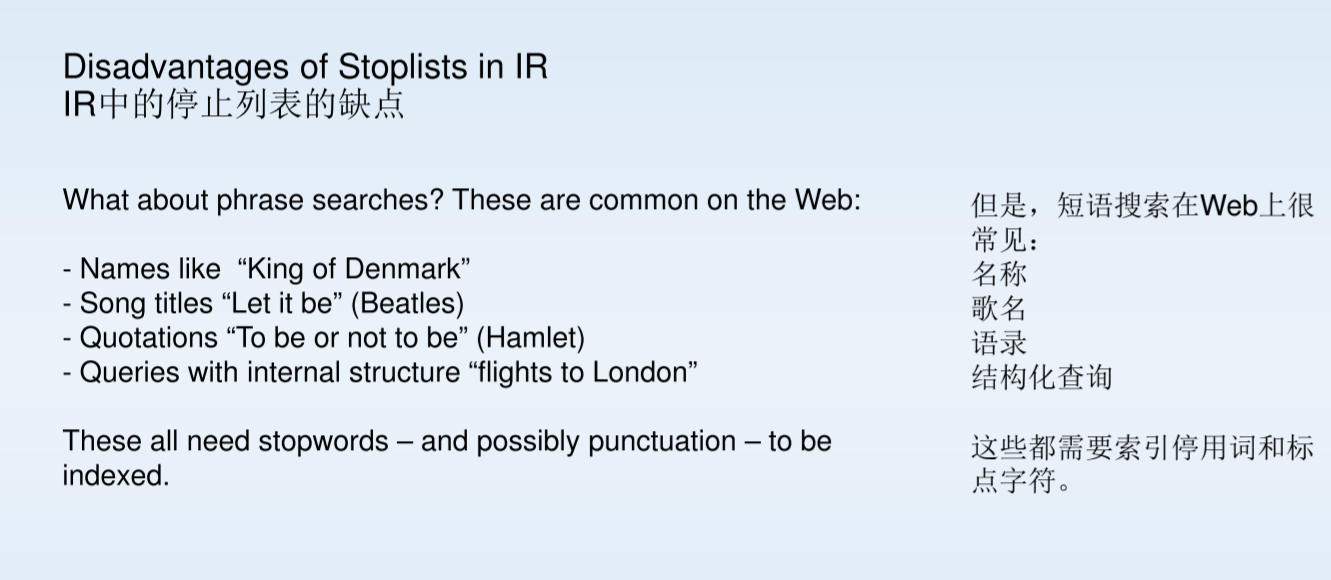
Questions could be on the following topics:

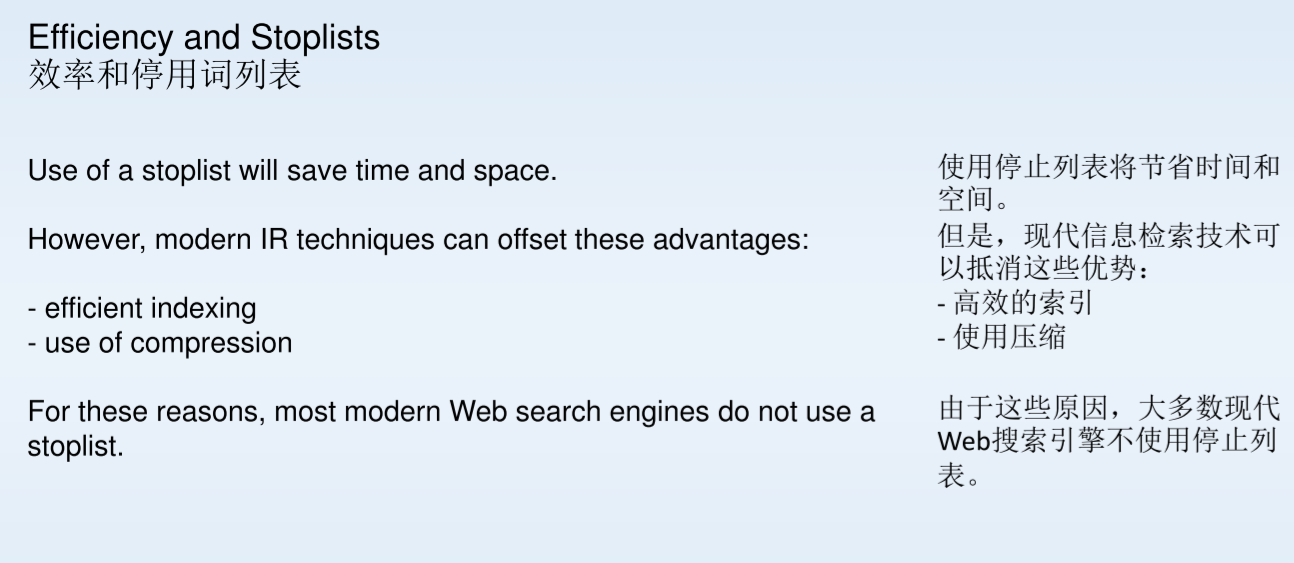
1.Luhn's graph and the work of Kucera and Francis. You need to understand in detail all the points that Luhn's graph shows. You need to understand the table of word frequencies taken from Kucera and Francis. You need to know how the Luhn graph and Kucera & Francis table relate to stoplist design.

Answer：

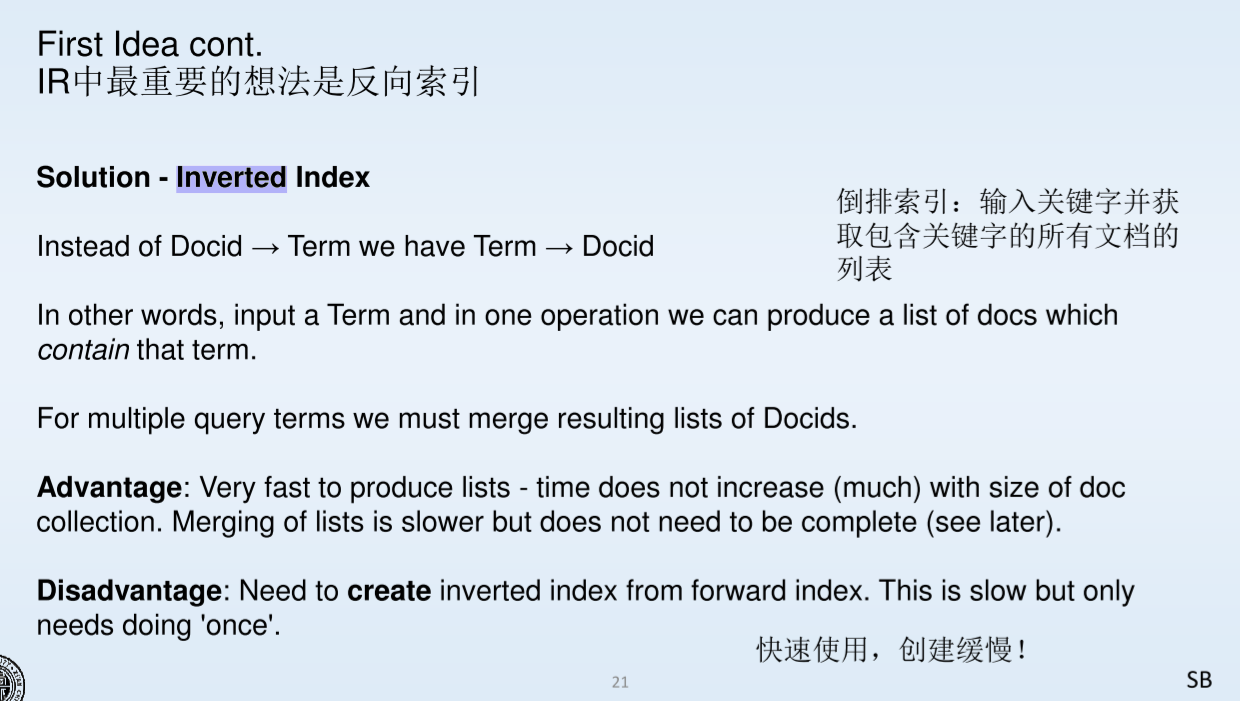
1. K and F created the Brown Corpus of Standard American English and conducted very important computational frequency studies on the corpus(语料库)
2. Luhn’s graph shows that freq of words falls off asymptotically - rapidly for first few words, then less and less rapidly as we move through the entire vocabulary.

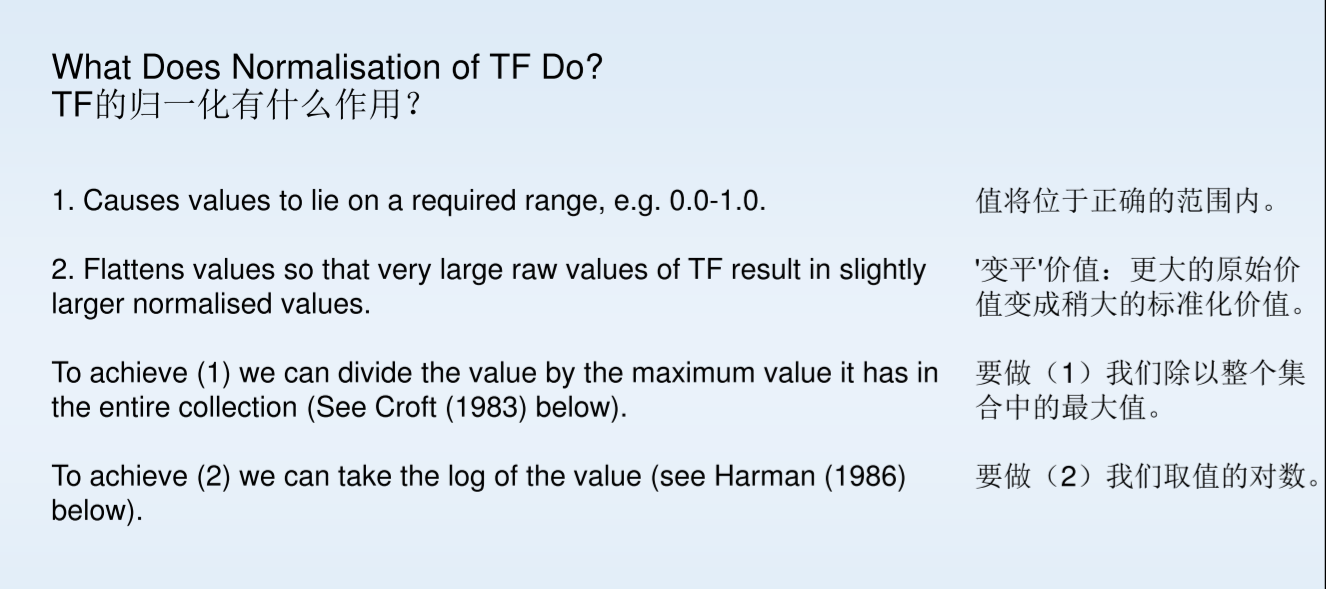


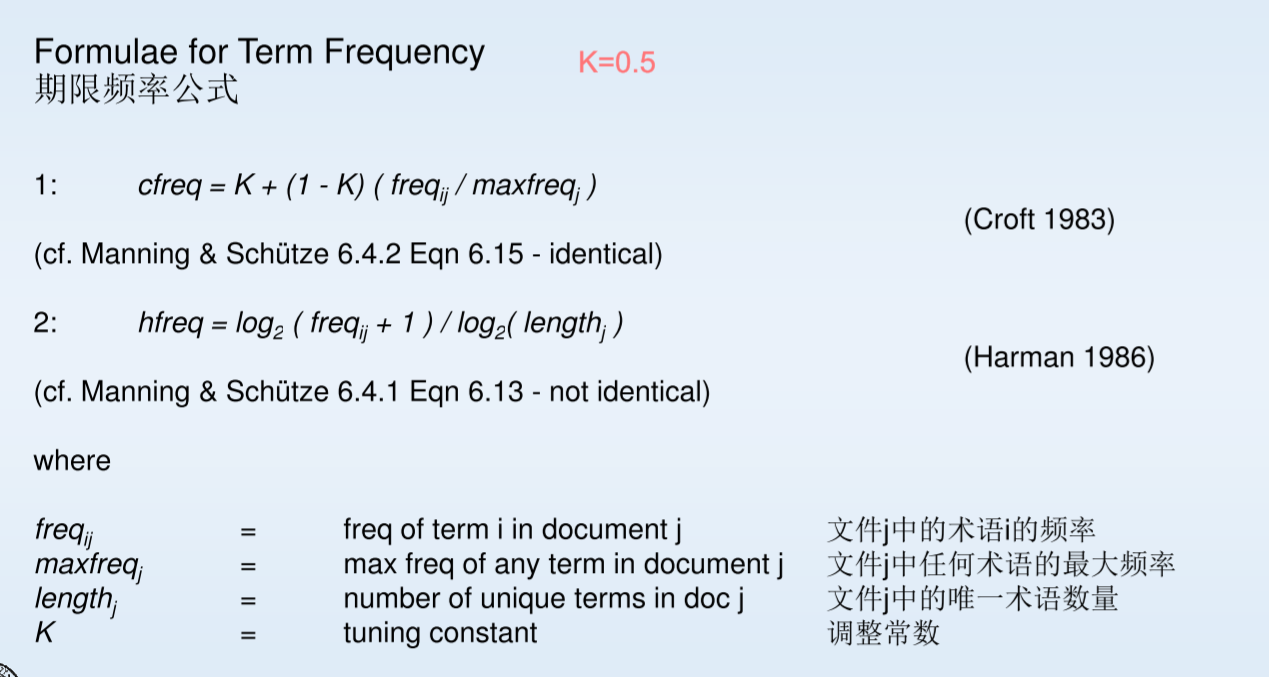
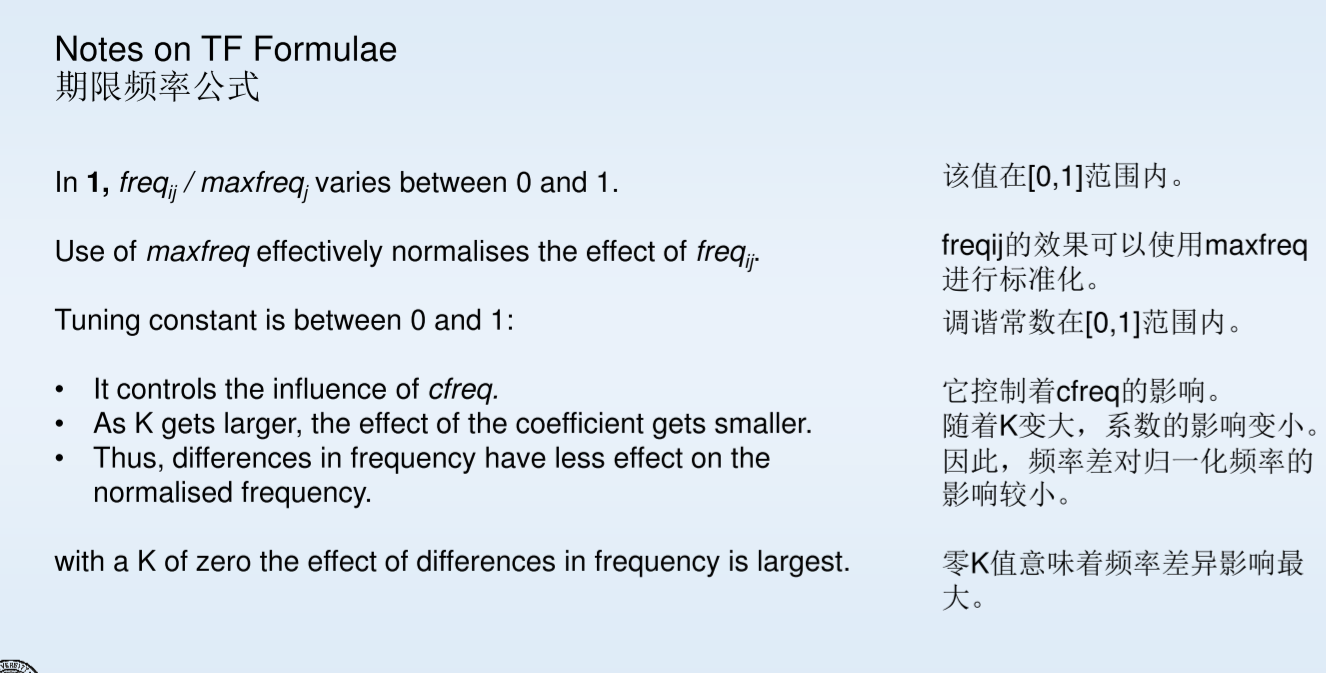
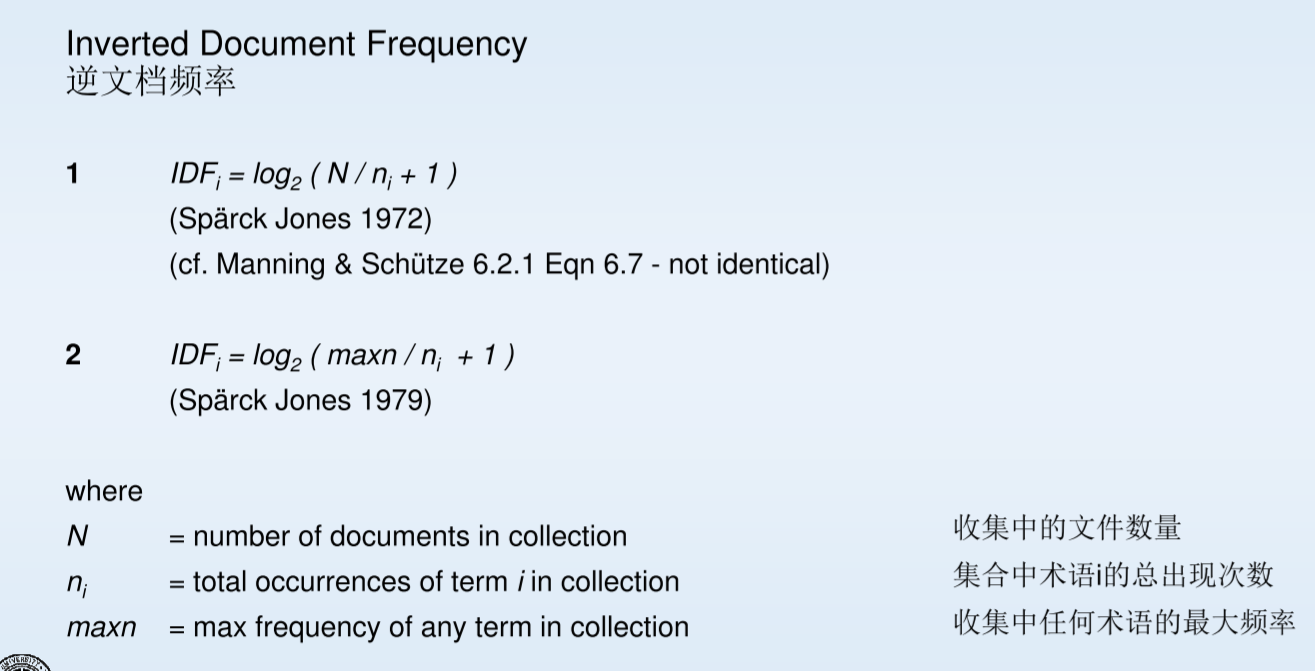


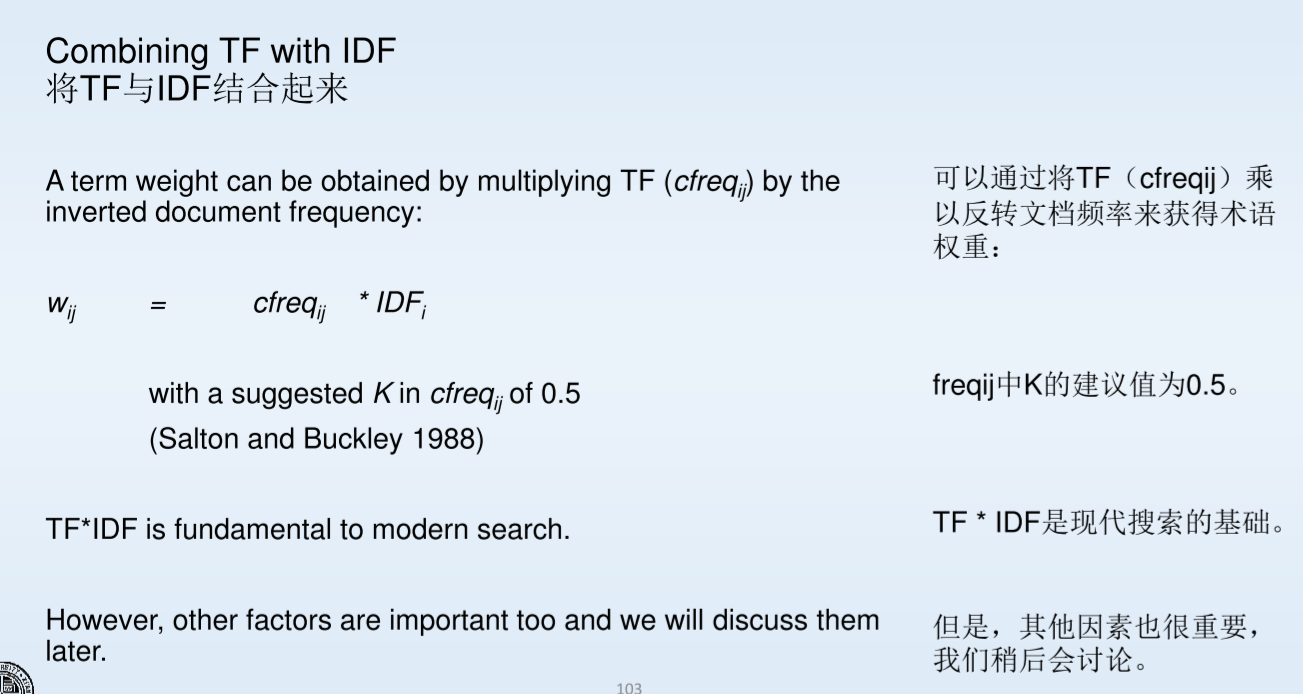


2.Inverted Index, Index File and Postings File. Also the computation of Boolean searches on Postings Lists. Searching the Index File using binary trees, in memory or on disk.



3.TF \* IDF model. How to compute TF and IDF using standard equations, based on example data. You would be provided with equations for TF and IDF so you do not need to memorise them. However, you do need to understand how to use them.





4.What is the purpose of TF and what is the purpose of IDF in IR? Make sure you really understand this.

(1) TF: This is a measure based on the frequency of the term within the document.

(2) IDF: inverted document freqency is a measure based on the frequency of the term within every document.

5.Evaluation and Ranking. You need to remember and understand the formulae for P, R and F (F1 version only, where P and R are equally weighted). You need to understand Precision at Rank, Average Precision, Mean Average Precision, Reciprocal Rank and Mean Reciprocal Rank. You could be asked to compute P and R for different levels of n, given some data.

6.PageRank. How to compute PageRank, based on an example network. The PageRank equation would be provided in such a question, but you need to know how to apply it. You need to know what the purpose of PageRank is, and how you could use it in IR. Query Dependent scores for a document and Query Independent scores for a document. What is the difference? What scores can be used? How to build into an IR system?

7.Web Crawling. You need to know an algorithm for Web Crawling and to be able to apply it to an example network. You need to be able to demonstrate your crawl of a network. See the notes on Web search.

8.Test Collections. You need to know methods of creating a test collection for IR. How to use a test collection for IR system evaluation.