Docs and Data

What are some problems with documents in IRs?

Docs and Data

What are the three classes of need that Janser et al. (2008) identified and what percentage of queries do they contain?

Docs and Data

What is usually contained in a navigational query?

Docs and Data

What is usually contained in a transactional query?

Docs and Data

What is usually contained in an informational query?

Docs and Data

Define 'index language'

Docs and Data

What is a document-term matrix?

Docs and Data

What do we do if we want to perform the query 'Brutus AND Caesar' in the prescence of a document-term matrix?

- Informational (∼80%)
- Navigational (∼10%)
- Transactional (∼10%)

- Expressed in natural language
- Info we seek may not be in out language
- Language is ambiguous
- No. of documents rapidly increasing

- Terms related to movies, songs, recipes etc.
- Queries with 'obtaining' terms.
- Download terms.
- Entertainment terms.
- Interaction terms ('buy', 'chat').

- Company/business/organisation name.
- Domain suffix (.com, .co.uk).
- Contains 'web'.
- Query is less than 3 words.

A means of representing documents and of representing queries.

Use of question words.
Many phrases.
Informational items e.g. 'list of'.
Query length larger than 2 words.

Take vectors for each term and perform a bitwise AND.

A mathematic matrix that describes the frequency of terms that occur in a collection of documents.

Docs and Data

Given the following document-term matrix, answer the query 'Brutus AND Caesar'

	Antony & Cleopatra	*Julius Caesar*	*The Tempest*
Antony	*1*	*0*	*0*
Brutus	*1*	*1*	*0*
Caesar	*1*	*1*	*1*

Docs and Data

What is the boolean retrieval model?

Docs and Data

What is the big problem with document-term matrices?

Docs and Data

What is the solution to having huge document-term matrices that have a huge amount of 1s and 0s?

Docs and Data

What is the concept of an inverted index?

Docs and Data

What is the name of a term in an inverted index?

Docs and Data

What are the four major steps in building an index?

Docs and Data

In what form is the input to an inverted indexer?

A model for information retrieval in which we can pose any query which is in the form of a boolean expression of terms.	110 & 111 = 110 Antony & Cleopatra, Julius Caesar
Only store the 1s which leads to using an inverted index.	If we have a realistic collection of documents ($N = 1M$) in which each document hold 1,000 terms. If we assume 500,000 unique words in the whole collection, our matrix will have half a trillion 1s and 0s.
A posting	1. Keep a dictionary of terms. 2. For each term, have a list that records which documents it occurs in.
A sequence of (modified token, document ID) pairs.	1. Collect documents to be indexed. 2. Tokenize text, turning each document into a list of tokens (tokenizer). 3. Do linguistic preprocessing, producing normalized tokens, which are the indexing terms (linguistic modules). 4. Index the documents that each term occurs in by creating an inverted index (indexer)

Docs and Data

What happens straight after an inverted indexer receives its input?

Docs and Data

What happens straight after an inverted indexer's input is sorted?

Docs and Data

How is a boolean AND query processed by using an inverted index?

Docs and Data

How can AND queries be optimized when using an inverted index?

For example: Brutus AND Caesar AND Caplurnia

Docs and Data

How can OR/AND queries be optimized when using an inverted index?

Docs and Data

Given the following, what's the processing order?

Term	*Frequency*
eyes	*213,312*
cat	*87,009*
dog	*107,913*
skies	*271,658*
owl	*46,653*
trees	*316,812*

(cat OR trees) & (dog OR skies) & (owl OR eyes)

Docs and Data

What are some Boolean model pros?

Docs and Data

What are some Boolean model cons?

Multiple term entries in a single document are merged and then data is split into dictionary and postings list.

The input is sorted by modified token and then by document ID.

Process the query in order of increasing frequency i.e. process the query with the smallest summed postings list size first.

1. Locate the terms in the index.
2. Intersect their postings lists.

1. (dog OR skies) AND (owl OR eyes)
2. (1) AND (cat OR trees)

Estimate the size of each OR query by summing the document frequencies of each term, then process by lowest frequency first.

- AND gives too few or no results, OR gives too many.
- No ordering of results.
- Order of words in document not used.
- Basic boolean expressions too limiting for informational needs.
- Non-expert user doesn't understand boolean operators.

- Simple model, easy/efficient to implement.
- Precise. Document either matches or doesn't.
- Widely used for commercial, legal retrieval and for specialist searches.

Docs and Data

What are the three main classes of tokens involved in tokenization?

Docs and Data

What is a major issue with splitting documents on spaces?

Docs and Data

In IR, what is a stop word?

Docs and Data

Why are stop words kept in indexes nowadays?

Docs and Data

In IR, what is token normalization?

Docs and Data

In IR, what is stemming?

Docs and Data

What are some issues that come with stemming?

Docs and Data

What is the name of a popular stemming algorithm?

Some languages don't use spaces to distinguish a split in words such as Chinese.	• Morphosyntactic word • Punctuation mark of special symbol • A number
Because they can add context to searches and the cost of storing and processing them is constantly reducing.	A stop word is a highly frequently occuring word which is filtered out as it has low distinguishing power.
The process of chopping 'ends of words' before indexing.	The process of mapping tokens to their normalised form e.g. B.B.C $\rightarrow$ BBC
Porter's stemming algorithm.	• It is a crude process and may yield forms that are 'not words'. • Under-stemming fails to conflate related forms. • Over-stemming conflates unrelated forms.

Docs and Data

What are the two measures used to evaluate the performance of search engines and what do they mean? How do they relate?

Docs and Data

What does a term-document matrix show?

Docs and Data

What is term frequency?

Docs and Data

What is the document frequency of term t?

Docs and Data

What is the formula for inverse document frequency of a term t?

Docs and Data

Why is log used in the calculation of a terms inverse document frequency?

Docs and Data

Docs and Data

What is the formula for the tf-idf weighting of a term, t, in a document, d?

How many times a word occurs in a document.

- Precision - the fraction of retrieved documents that are relevant.
- Recall - the fraction of relevant documents that are retrieved.
- Precision increases are recall decreases.

The number of documents that contains the term t.

The amount of times a term t appears in a document d.

It dampens the effect of idf.

$$idf_t = log \frac{N}{df_t}$$

N: number of documents in collection.
df_t: document frequency of term t.

$$W_{t,d} = (1 + log(tf_{t,d})) * log_{10}\left(\frac{N}{df_t}\right)$$

Docs and Data

How do we calculate the final ranking score of a document given a query?

Docs and Data

How can vectors be used to rank documents given a query?

Docs and Data

Why shouldn't we measure vector space proximity in terms of Euclidean distance?

Docs and Data

What is the alternative to measuring vector space proximity with Euclidean distance?

Docs and Data

What is the process of measuring the angle between two vectors?

Docs and Data

What is the angle between the vectors (7, 10) and (13, 2)?

Docs and Data

What does the PageRank algorithm do?

Docs and Data

What notion is used in the PageRank algorithm?

Represent both documents and query as vectors and rank the documents according to their proximity to the query in the vector space.	$\text{Score}_{q,d} = \sum_{t \in q \cap d} \text{tf.idf}_{t,d}$ Add up tf-idf weight of each query term found in d.
Measuring the angle between vectors.	Because Euclidean distance can be very large for vectors of different length although they are close to each other in the space.
$\theta \approx 46.3$	1. Length-normalize the vectosrs (divide components by magnitude). 2. Calculate the dot product of the normalized vectors. 3. Inverse cosine the result.
A random web surfer is used who goes from the current page to a random one that the current page links to. Another option is that the surfer teleports to a random page that the current page doesn't link to.	Calculates a score for each page by using the hyperlinked structure of the web.

Docs and Data

How does the PageRank teleport operation work?

Docs and Data

What is the equation to calculate the PageRank score of a page, x?

Docs and Data

Docs and Data

Docs and Data

Docs and Data

Docs and Data

Docs and Data

C_t: out-degree of t

$$PR(x) = \alpha\left(\frac{1}{N}\right) + (1\alpha)\sum_{i=1}^{n}\frac{PR(t_i)}{C(t_i)}$$

Left side of +: probability of a random jump
Right side of +: probability of a normal click multiplied by the sum of the contributions of all pages contributing IN links to x based on the number of OUT links for each such page and its own PageRank score.

- When at a node with no out-links, invoke the teleport operation.
- When at a node with out-links, invoke the teleport operation with probability $0 < x < 1$. Usually x is around 15.

Docs and Data	Docs and Data
Docs and Data	Docs and Data
Docs and Data	Docs and Data
Docs and Data	Docs and Data

Docs and Data	Docs and Data
Docs and Data	Docs and Data
Docs and Data	Docs and Data
Docs and Data	Docs and Data

Docs and Data	Docs and Data
Docs and Data	Docs and Data
Docs and Data	Docs and Data
Docs and Data	Docs and Data

Docs and Data	Docs and Data
Docs and Data	Docs and Data
Docs and Data	Docs and Data
Docs and Data	Docs and Data

Docs and Data	Docs and Data
Docs and Data	Docs and Data
Docs and Data	Docs and Data
Docs and Data	Docs and Data

Docs and Data

Docs and Data

Docs and Data