Information Retrieval System

Shuai Wei
Zian Chen
Yunqing Zhang

Outline

Implement Comparison

Comparison of our algorithm in boolean and tf-idf

Performance Analysis

Performance analysis of Boolean, tf-idf and BM25

New Idea

Introduction of a new IR model – ’DFR’

01

Implementation Comparison

Result Compare of tf-idf

Search “World cup champion”

Shuai Wei	Zian Chen	Yunjing Zhang
1.retrieval text part without title 2.didn’t compute idf in document,only in term	1.retrieval the whole document including title	1.retrieval text part without title

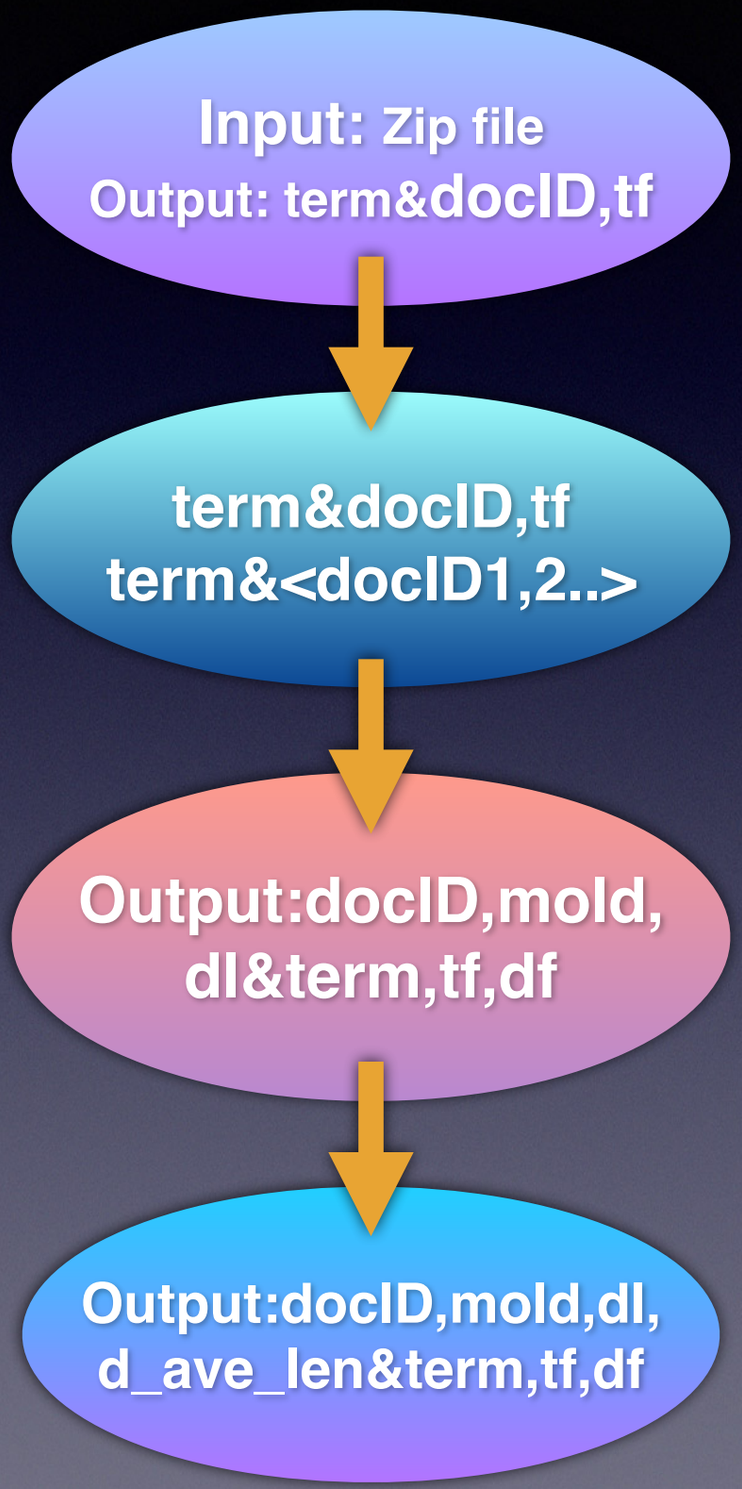
Result Compare of precision&recall

Shuai Wei	Zian Chen	Yunjing Zhang
1.precision:64% 2.recall:73%	1.precision:56% 2.recall:76%	1.precision:63% 2.recall:68%

02

Performance Analysis

BM25 Implementation



BM25 Implementation

Find “relevant” documents

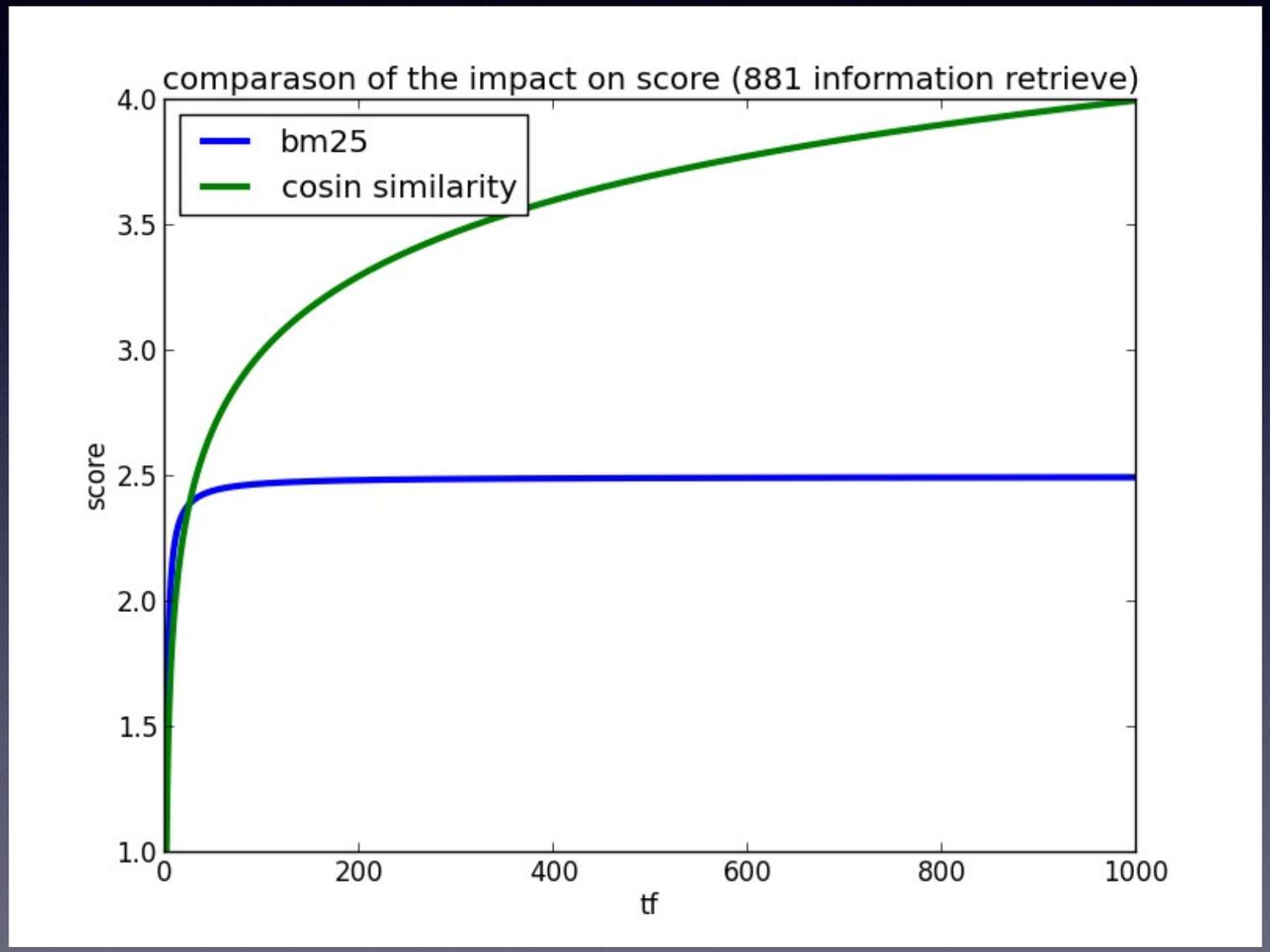
Compute score for relevant docs

$$RSV_d = \sum_{t \in q} \log\left[\frac{N}{df_t}\right] \cdot \frac{(k_1 + 1)tf_{td}}{k_1((1-b) + b \cdot (L_d / L_{ave})) + tf_{td}}$$

Tree map to rank top 10

Retrieve content of top 10

tf saturation for tf-idf and BM25



Background

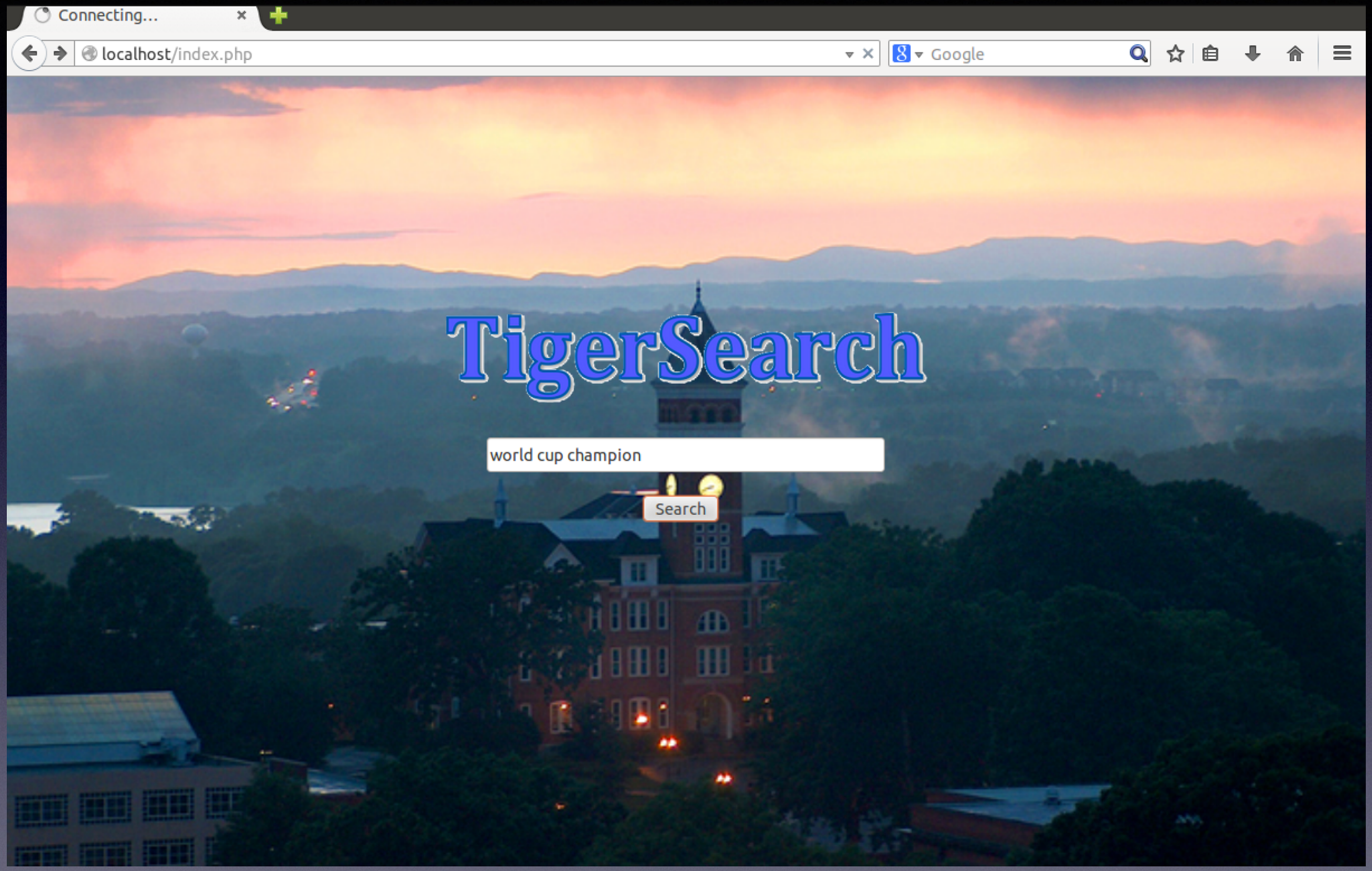
Techniques we use

HTML,CSS,PHP

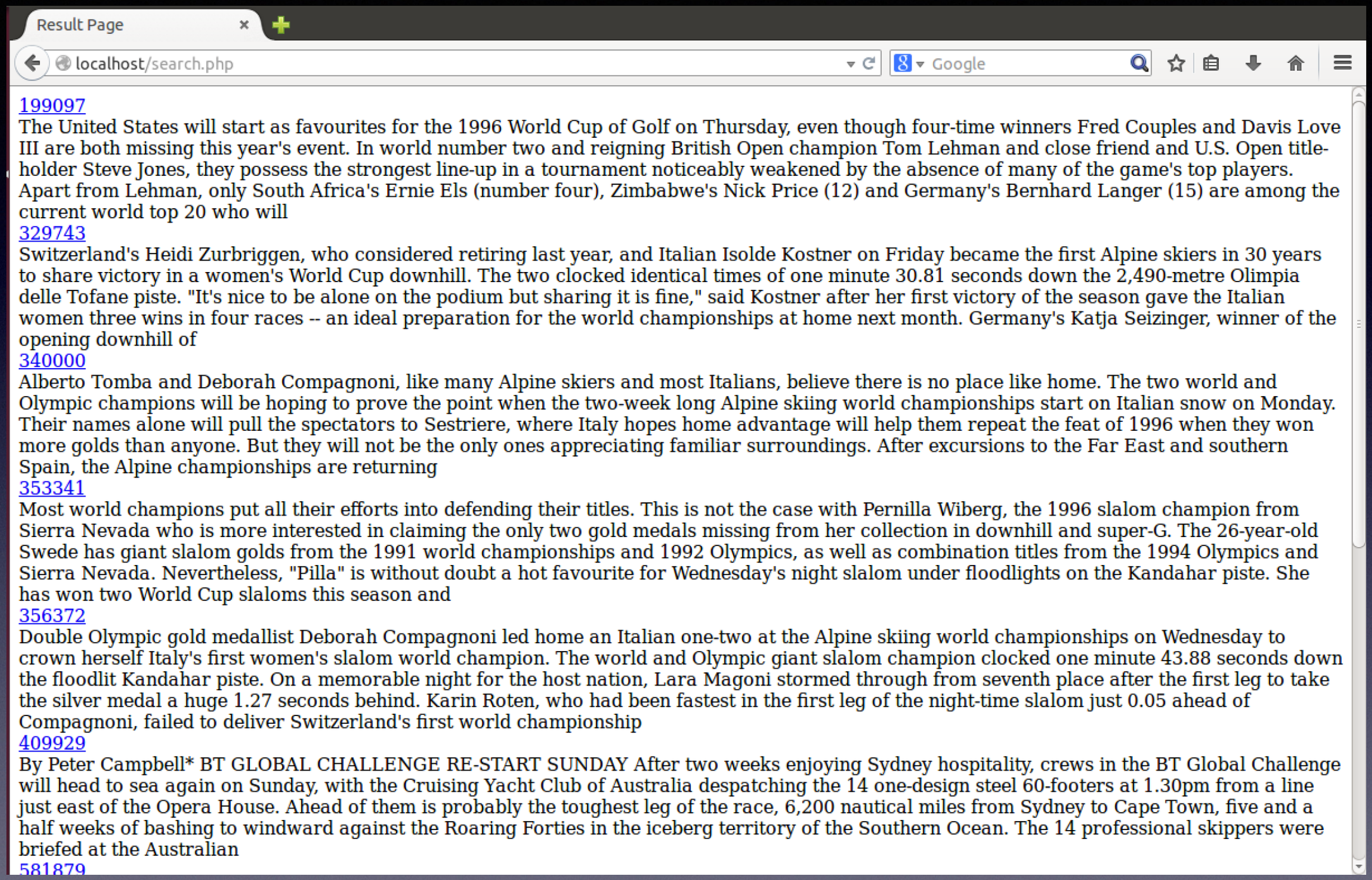
Hadoop

SSH2

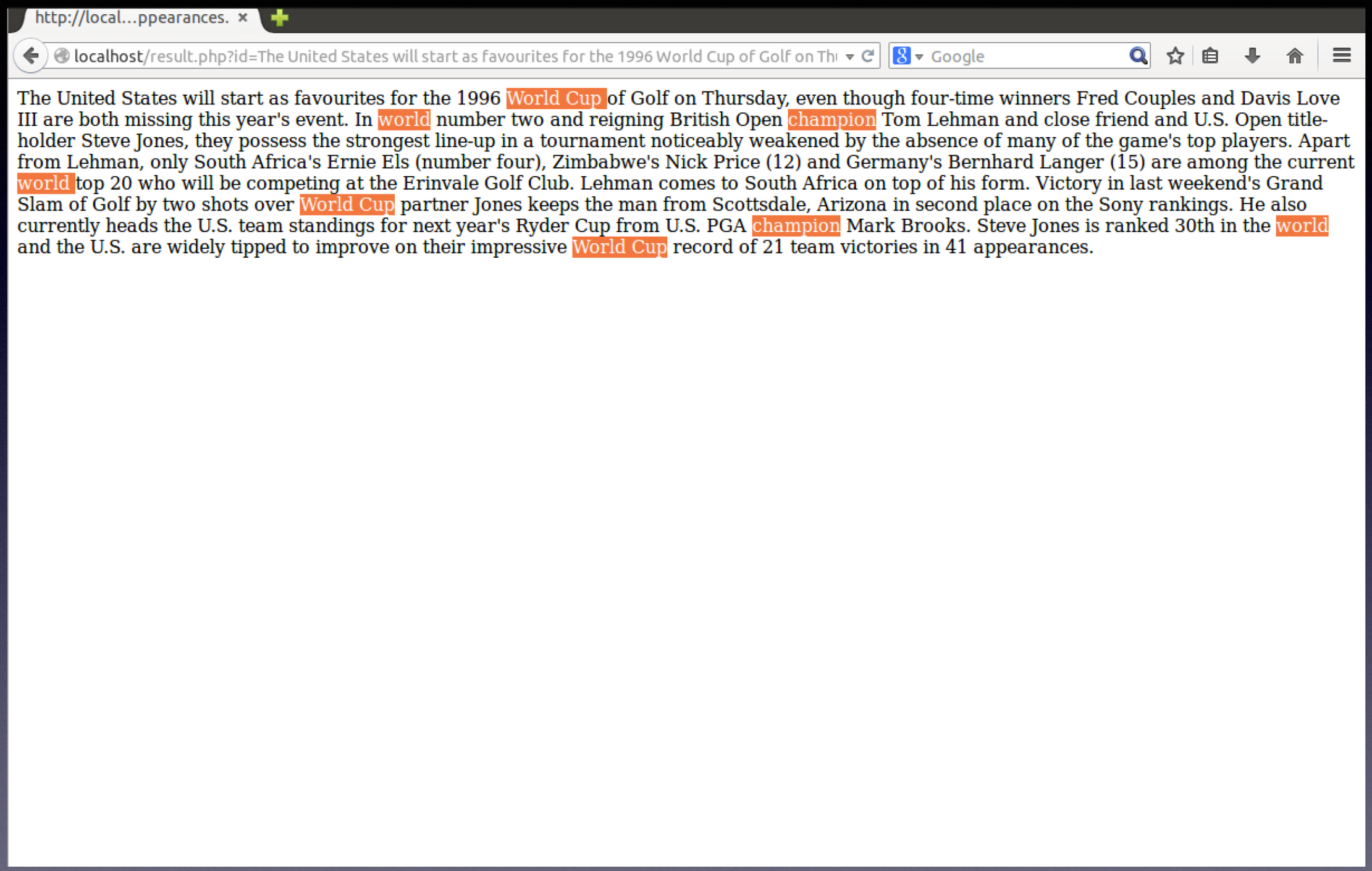
Demo


Connecting...
localhost/index.php
Google
TigerSearch
world cup champion
Search

Demo


Result Page
localhost/search.php
Google
199097
The United States will start as favourites for the 1996 World Cup of Golf on Thursday, even though four-time winners Fred Couples and Davis Love III are both missing this year's event. In world number two and reigning British Open champion Tom Lehman and close friend and U.S. Open title-holder Steve Jones, they possess the strongest line-up in a tournament noticeably weakened by the absence of many of the game's top players. Apart from Lehman, only South Africa's Ernie Els (number four), Zimbabwe's Nick Price (12) and Germany's Bernhard Langer (15) are among the current world top 20 who will
329743
Switzerland's Heidi Zurbriggen, who considered retiring last year, and Italian Isolde Kostner on Friday became the first Alpine skiers in 30 years to share victory in a women's World Cup downhill. The two clocked identical times of one minute 30.81 seconds down the 2,490-metre Olimpia delle Tofane piste. "It's nice to be alone on the podium but sharing it is fine," said Kostner after her first victory of the season gave the Italian women three wins in four races -- an ideal preparation for the world championships at home next month. Germany's Katja Seizinger, winner of the opening downhill of
340000
Alberto Tomba and Deborah Compagnoni, like many Alpine skiers and most Italians, believe there is no place like home. The two world and Olympic champions will be hoping to prove the point when the two-week long Alpine skiing world championships start on Italian snow on Monday. Their names alone will pull the spectators to Sestriere, where Italy hopes home advantage will help them repeat the feat of 1996 when they won more golds than anyone. But they will not be the only ones appreciating familiar surroundings. After excursions to the Far East and southern Spain, the Alpine championships are returning
353341
Most world champions put all their efforts into defending their titles. This is not the case with Pernilla Wiberg, the 1996 slalom champion from Sierra Nevada who is more interested in claiming the only two gold medals missing from her collection in downhill and super-G. The 26-year-old Swede has giant slalom golds from the 1991 world championships and 1992 Olympics, as well as combination titles from the 1994 Olympics and Sierra Nevada. Nevertheless, "Pilla" is without doubt a hot favourite for Wednesday's night slalom under floodlights on the Kandahar piste. She has won two World Cup slaloms this season and
356372
Double Olympic gold medallist Deborah Compagnoni led home an Italian one-two at the Alpine skiing world championships on Wednesday to crown herself Italy's first women's slalom world champion. The world and Olympic giant slalom champion clocked one minute 43.88 seconds down the floodlit Kandahar piste. On a memorable night for the host nation, Lara Magoni stormed through from seventh place after the first leg to take the silver medal a huge 1.27 seconds behind. Karin Roten, who had been fastest in the first leg of the night-time slalom just 0.05 ahead of Compagnoni, failed to deliver Switzerland's first world championship
409929
By Peter Campbell* BT GLOBAL CHALLENGE RE-START SUNDAY After two weeks enjoying Sydney hospitality, crews in the BT Global Challenge will head to sea again on Sunday, with the Cruising Yacht Club of Australia despatching the 14 one-design steel 60-footers at 1.30pm from a line just east of the Opera House. Ahead of them is probably the toughest leg of the race, 6,200 nautical miles from Sydney to Cape Town, five and a half weeks of bashing to windward against the Roaring Forties in the iceberg territory of the Southern Ocean. The 14 professional skippers were briefed at the Australian
581879

Demo


http://local...ppearances.
localhost/result.php?id=The United States will start as favourites for the 1996 World Cup of Golf on Th
Google
The United States will start as favourites for the 1996 World Cup of Golf on Thursday, even though four-time winners Fred Couples and Davis Love III are both missing this year's event. In world number two and reigning British Open champion Tom Lehman and close friend and U.S. Open title-holder Steve Jones, they possess the strongest line-up in a tournament noticeably weakened by the absence of many of the game's top players. Apart from Lehman, only South Africa's Ernie Els (number four), Zimbabwe's Nick Price (12) and Germany's Bernhard Langer (15) are among the current world top 20 who will be competing at the Erinvale Golf Club. Lehman comes to South Africa on top of his form. Victory in last weekend's Grand Slam of Golf by two shots over World Cup partner Jones keeps the man from Scottsdale, Arizona in second place on the Sony rankings. He also currently heads the U.S. team standings for next year's Ryder Cup from U.S. PGA champion Mark Brooks. Steve Jones is ranked 30th in the world and the U.S. are widely tipped to improve on their impressive World Cup record of 21 team victories in 41 appearances.

BM25 vs. tf-idf

	tf-idf	BM25
Origins from	Vector space model	probabilistic relevance model
performance for short docs	good	better
performance for long docs	better	good
two params affect if saturation	don’t have	have
Similar	• Use inverse document frequency to distinguish between common (low value) words and uncommon (high value) words. • Both recognize (see Term frequency) that the more often a word appears in a document, the more likely is it that the document is relevant for that word.	

Result Compare

Search “Nelson Mandela and senior African National Congress associates were jailed for life for resisting apartheid”

tf-idf	BM25
4966 46.29155098531942	5914 0.30306876105760827
5849 46.69070372609274	5905 0.30338657591718887
5845 47.19199143958318	708457 0.30629735444715894
3175 47.747772226276055	437258 0.3072169332150732
5914 51.006698809345394	99392 0.3190984261320096
5910 52.0230374272586	555486 0.31949260912799343
3112 54.16731386155291	555492 0.330326531967224
5847 54.57505366326651	594346 0.34332221423493886
3116 61.0883221309362	24201 0.35003248515266555
5905 85.39512848684674	601467 0.3718526900547791

Pros. and Cons. of tf-idf

Efficient and simple algorithm for matching words in a query to document that are relevant to that query

tf-idf returns documents that are highly relevant to a particular query

Compute doc similarity in word count space

It makes no use of semantic similarity between words
tf-idf does not make the jump to the relationship between words

Pros. and Cons. of Boolean

Easy to implement and it is computationally

enables users to express structural and conceptual constraints to describe important linguistic features

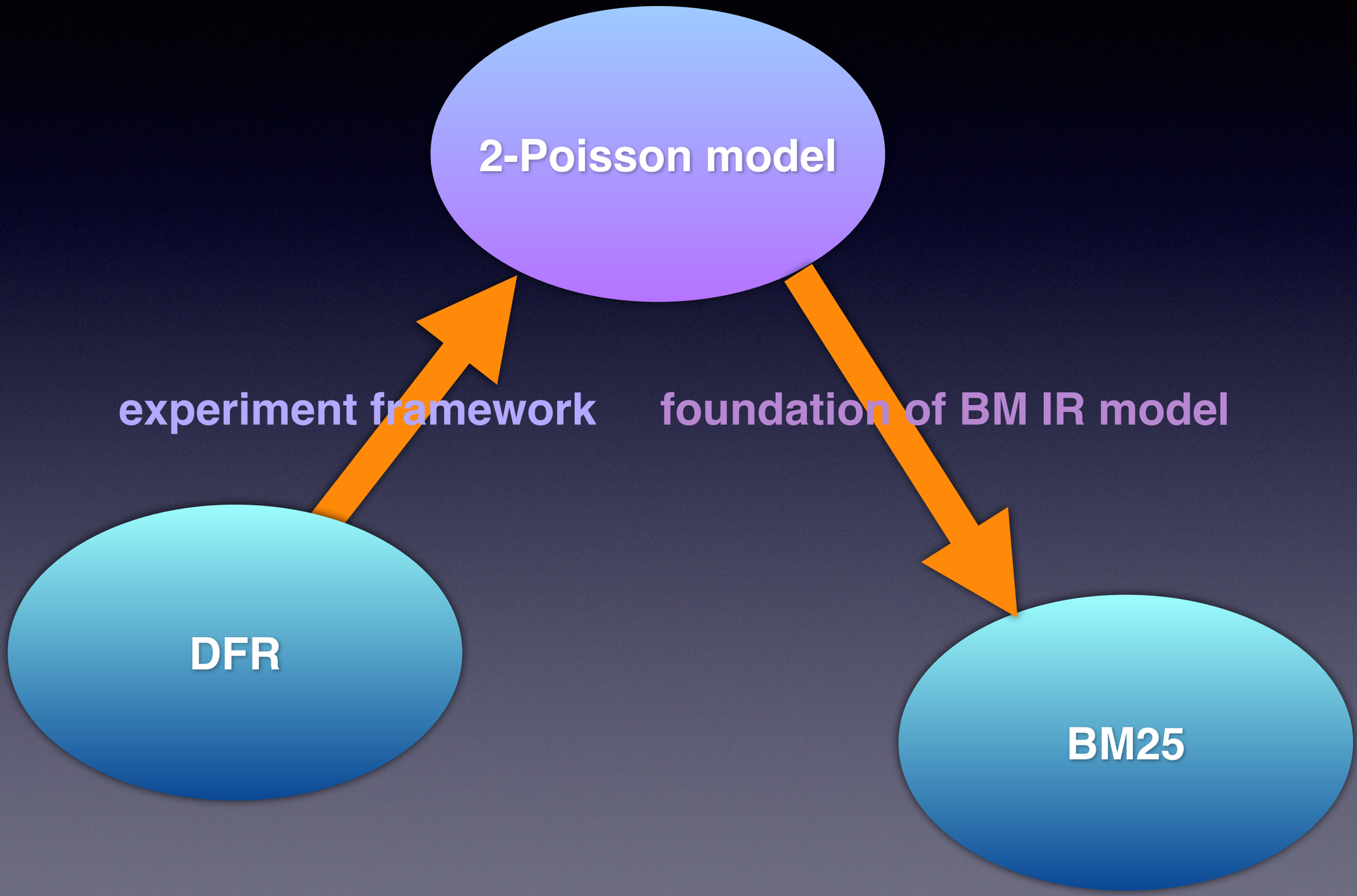
Only documents that satisfy a query exactly are retrieved.

Difficult to control the number of retrieved documents. Traditional Boolean approach does not provide a relevance ranking of the retrieved documents,

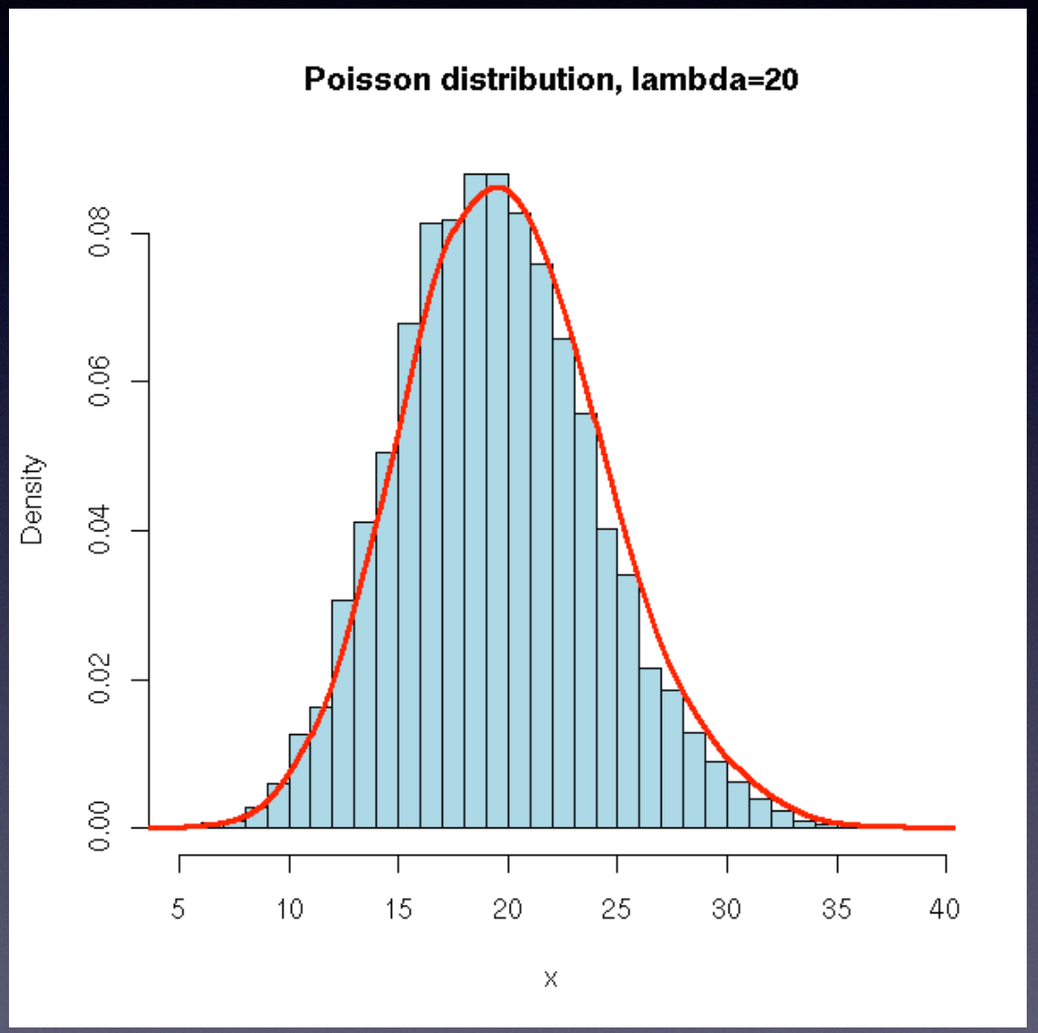
03

New Idea

Divergence From Randomness



2 Poisson Model



Proposed Method

Select a basic randomness model

- binomial
- geometric
- Term frequency

↓

applying the first normalization

↓

Normalizing the term frequencies

Basic Randomness Model

The more the divergence of the within-document term-frequency from its frequency within the collection, the more the information carried by the word t in the document d

$$weight(t \mid d) \propto -\log \Pr ob_M(t \in d \mid Collection)$$

M: the type of model of randomness

Basic DFR Models	
D	Divergence approximation of the binomial
P	Approximation of the binomial
B_F	Bose-Einstein distribution
G	Geometric approximation of the Bose-
I(n)	Inverse Document Frequency model
I(F)	Inverse Term Frequency model
$I(n_e)$	Inverse Expected Document Frequency

Basic Randomness Model

Example of M model (binomial)

$$-\log \Pr ob_M(t \in d \mid Collection) = -\log \left(\frac{TF}{tf} \right) p^{tf} q^{TF-tf}$$

TF is the term-frequency of the term t in the collection
tf is the term-frequency of the term t in the document d
N is the number of document in the collection
p is 1/N and q=1-p

First Normalization

However, we need to import P_{risk} to smooth the influence of term frequency on Probability

$$gain(t \mid d) = P_{risk} \bullet (-\log \Pr ob_M(t \in d \mid Collection))$$

$$P_{risk} = 1 - \Pr ob(t \in d \mid d \in Eliteset))$$

Term Frequency Normalization

Before using Formula(2), we need to normalize tf and document length(dl)

$$tfn = tf \bullet \log(1 + \frac{sl}{dl})$$