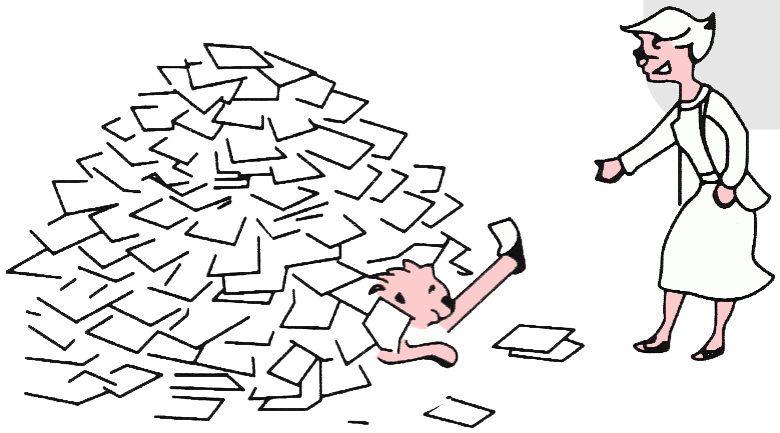
https://vvtesh.sarahah.com/

Information Retrieval

Venkatesh Vinayakarao

Term: Aug – Dec, 2018
Indian Institute of Information Technology, Sri City



The **Google effect**, also called digital amnesia, is the tendency to forget information that can be found readily online by using Internet search engines such as **Google**. According to the first study about the **Google effect** people are less likely to remember certain details they believe will be accessible online.
-Wikipedia.

Instructions

INFORMATION RETRIEVAL — MOCK MID-TERM 1

Roll Number: _____ TOTAL = 20 MARKS. TIME = 90 MINUTES.
Name: _____

Instructions

- This is a closed book test. You are not allowed to carry any printed material with you.
- You are allowed to carry one page (A4 or smaller) hand written notes. Write your name and roll number on these notes.
- Please switch off your mobile phones and any other digital equipment you may have (like smart watches, calculators).
- Wherever you see <yourrollno>, replace this text with your 12 char/digit roll number.
- A negative mark of -2 applies to all questions.
- Write all answers correct up to two decimal places.
- Put down your final answer in this sheet. Give a succinct explanation for your answer. Give your explanation in a separate sheet. Explanation will not be graded.

Stemmer

http://snowball.tartarus.org/download.html

Introduction
Demo
Download
Mailing lists
License
Credits
Projects

Source on github

Snowball - Download

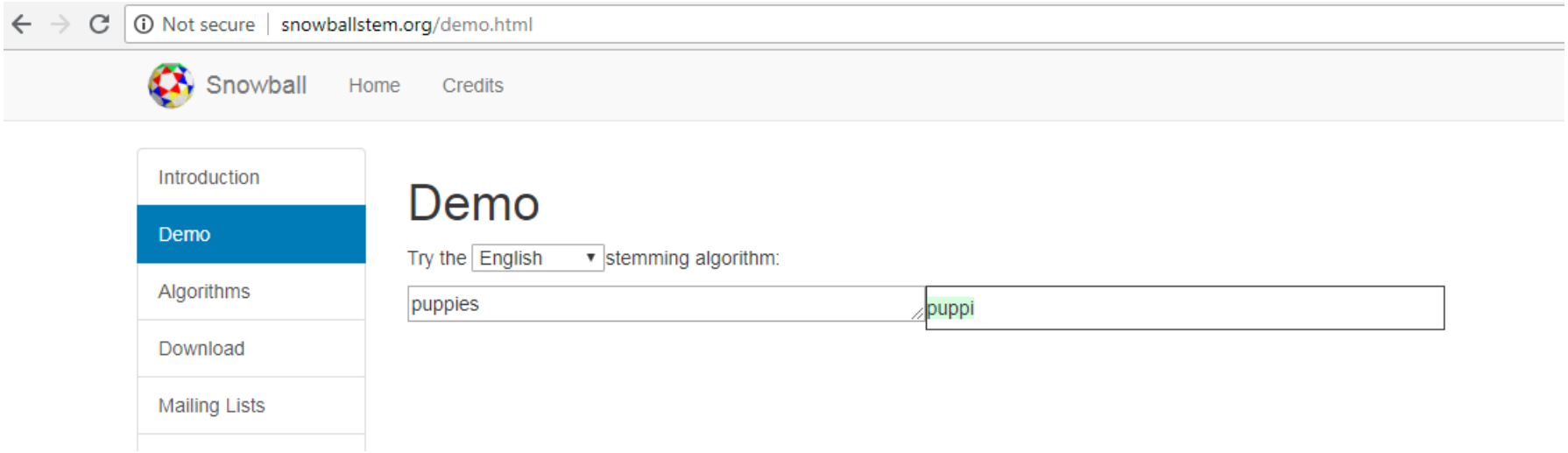
Tarballs

Several tarballs of the snowball sources are available.

- The C version of the libstemmer library.
  This contains all you need to include the snowball stemming algorithms into a C project of your own. If you downl
  or worry about the internals of the stemmers in any way.
- The Java version of the libstemmer library.
  This contains all you need to include the snowball stemming algorithms into a Java project of your own. If you dov
  compiler, or worry about the internals of the stemmers in any way.
- Snowball, algorithms, and libstemmer library.
  This contains all the source code for snowball (but not the generated source files). This is useful mainly if you are

Demo

http://snowballstem.org/demo.html



Using Stemmers

```
1  package org.tartarus.snowball;
2
3  import org.tartarus.snowball.ext.englishStemmer;
4
5  public class MyStemmer {
6      public static void main(String[] args) {
7
8          SnowballStemmer stemmer = new englishStemmer();
9          stemmer.setCurrent("puppies");
10         stemmer.stem();
11         System.out.println(stemmer.getCurrent());
12     }
13 }
```

How do these Stemmers Work?

- Porters Algorithm
  - http://snowball.tartarus.org/algorithms/porter/stemmer.html
- Suffix Removal Rules

SSES	->	SS
IES	->	I
SS	->	SS
S	->	

Word Measures

- (m>1) EMENT →
  - replac**ement** vs. c**ement**
- Removes the prefix only if it is long enough.

Computing Size

Revision

Quiz

If you have 800000 documents, how many bits are required to store the document identifiers (docID)?

No. of bits

bit1	bit2
0	0
0	1
1	0
1	1

Four numbers → 2 bits

bit1	bit2	bit3
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

Eight numbers → 3 bits

No. of bits

2^k numbers can be represented using k bits.

Document Identifiers

- d1 = 1
- d2 = 2
- d3 = 3
- d4 = 4
- d5 = 5
- …

Quiz

If you have 800000 documents, how many bits are required to store the document identifiers (docID) field?

k numbers can be represented using $\log_2 k$ bits

800000 documents require $\log_2 800000$ bits

(approx.) 20 bits

Quiz

- What is the difference between log, $\log_2$, and ln?

Quiz

- A collection has
  - 800000 documents,
  - 200 tokens per document and
  - six characters per token,
- Assume token → term conversion rate is 50%
- What is the size of the collection (in MB)?

Quiz

- A collection has
  - 800000 documents,
  - 200 tokens per document and
  - six characters per token,
- Assume token → term conversion rate is 50%
- What is the size of the collection (in MB)?

Answer: 800000 * 200 * 6 bytes = 960 MB