Latent Semantic Analysis (LSA)
(Example)

Martin Russell

School of Computer Science

February 2020

EESE BEng Final Year Project Specifications

- 112 report specifications

- Example:

Third Year BEng Final Year Design Project - 2003/2004
Project Title: Software Routines for the Testing and Diagnostics of Jaguar's Body System Network
Student Name: A Student
Supervisor: Clive Roberts
Project Specification:
Background. (Please include a general scene-setting overview of the project - targeted at the non-specialist)
The project will be to design and construct an SCP (Standard Corporate Protocol) test and diagnostic tool for Jaguar. SCP is a bespoke communication protocol similar in functionality to CAN. A network analysis tool, JEST (Jaguar Engineering SCP Tool), already exists and is used for the Ford SCP. However, JEST's functionality has not been incorporated into software based automated test and diagnostic routines. SCP is currently used as the 'body system network' on all Jaguar vehicle, the body system network provides control for locking, seat position, lighting control, etc.

Text retrieval C programs

- `index.c`
  - Computes the term-document index for a given corpus

  `Index fileList > index` *entry*

- `doc2vec.c`

  TF-IDF
  - Converts a document in a corpus into a <mark>document vector</mark>

- `stop.c`
  - Stopword removal

  before

- `porter-stemmer.c`
  - Implementation of the porter stemmer

Process

- Apply stop word removal

- Apply Porter stemmer

- Apply `doc2vec` to create the word-document matrix

- Apply MATLAB script lda.m
  - Reads word document matrix
  - Applies singular value decomposition
  - Analyses first three "topic" vectors

- Apply `index.c` to create term-document index and access word list
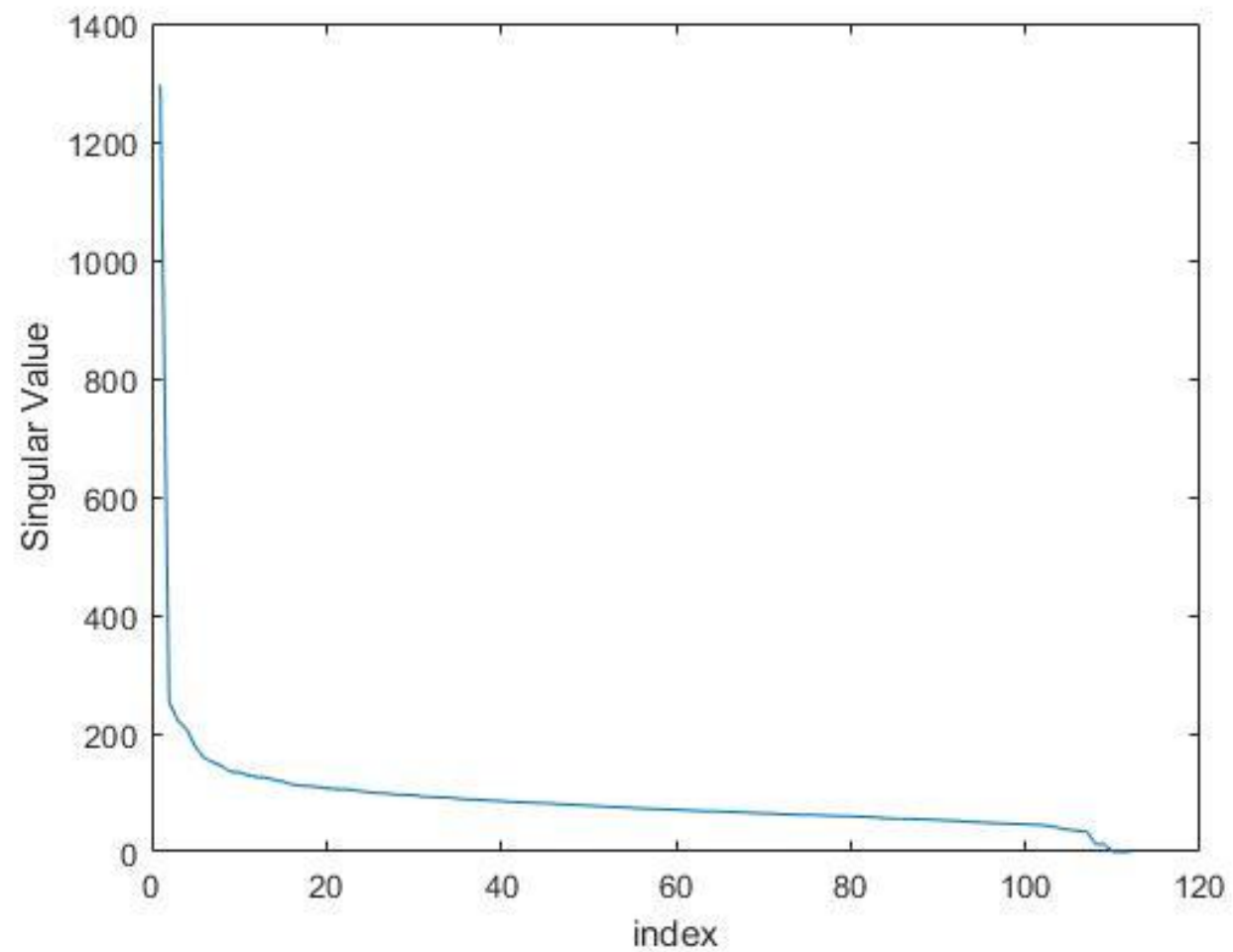
$\int \rightarrow$ text

LSA.m

```matlab
W = load('stemwddocmat');
[U,S,V]=svd(W);
% Extract singular values
voc=size(U,1);
sv=zeros(voc,1);
for i=1:1:voc
    sv(i)=S(i,i);
end
% Plot sigular values
x=1:1:voc;
plot(x,sv);
xlabel('index');
ylabel('Singular Value');
% Interpret sigular vector
sv1=V(:,1);
for i=1:1:10
    [argval,argmin]=min(sv1);
    argmin
    sv1(argmin)=0;
 end
…..
```

max V – column
         ↓
look into words

Singular values

Analysis of topic 1

1902 word=project wordCount=1576 docCount=111 IDF=0.008969
  1 docName=stem\AbassiM.stm count=14 weight=0.125561
  2 docName=stem\AgricoleW.stm count=14 weight=0.125561
  3 docName=stem\AliR.stm count=14 weight=0.125561
  4 docName=stem\AngCX.stm count=13 weight=0.116593
  5 docName=stem\AngeloZ.stm count=13 weight=0.116593
  6 docName=stem\AppadooD.stm count=14 weight=0.125561
  7 docName=stem\BenHasineA.stm count=13 weight=0.116593
  8 docName=stem\BenssiN.stm count=13 weight=0.116593
  9 docName=stem\BhagatR.stm count=15 weight=0.134530
  10 docName=stem\BradyE.stm count=13 weight=0.116593
  11 docName=stem\BriarsC.stm count=14 weight=0.125561
  12 docName=stem\Broadhurst.stm count=13 weight=0.116593
  13 docName=stem\BronksA.stm count=13 weight=0.116593
  14 docName=stem\BrownL.stm count=13 weight=0.116593

  ......

Analysis of topic 1

- Key terms:
  - Project
  - Outcom
  - Student
  - Pleas
  - Us
  - Expect
  - ….

General topic about projects?

Analysis of topic 2

- Data
- Reson
- Coupl
- Simul
- Expect
- Outcom
- Microstrip
- …

Microwave comms?

Analysis of topic 3

- Speech
- Variabl
- Should
- Recognit
- Signal
- …..

Speech recognition?

Summary

- Example application of LSA to a corpus
- C programs all available on Canvas
- Simple…