BLOCK POSTERS

Your settings for this Block Poster are:

Pages Wide 3

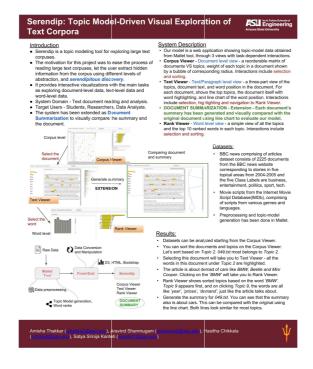
Orientation PORTRAIT

Paper Format A4

Border Setting Top: 0, Right: 0, Bottom: 0, Left: 0

You can find tips on printing, assembling and putting up your poster at www.blockposters.com

Enjoy your Block Poster!



Serendip: Topic Mode Text Corpora

Introduction

- Serendip is a topic modeling tool for e corpuses.
- The motivation for this project was to reading large text corpuses, let the us information from the corpus using diff abstraction, and serendipitous disce
- It provides interactive visualizations vas exploring document-level data, texword-level data.
- System Domain Text document read
- Target Users Students, Researcher
- The system has been extended as Description to visually compare the document.



el-Driven Visual Explora

exploring large text

ease the process of ser extract hidden erent levels of

overy.

vith the main tasks ct-level data and

ding and analysis. s, Data Analysts.

ocument

the summary and



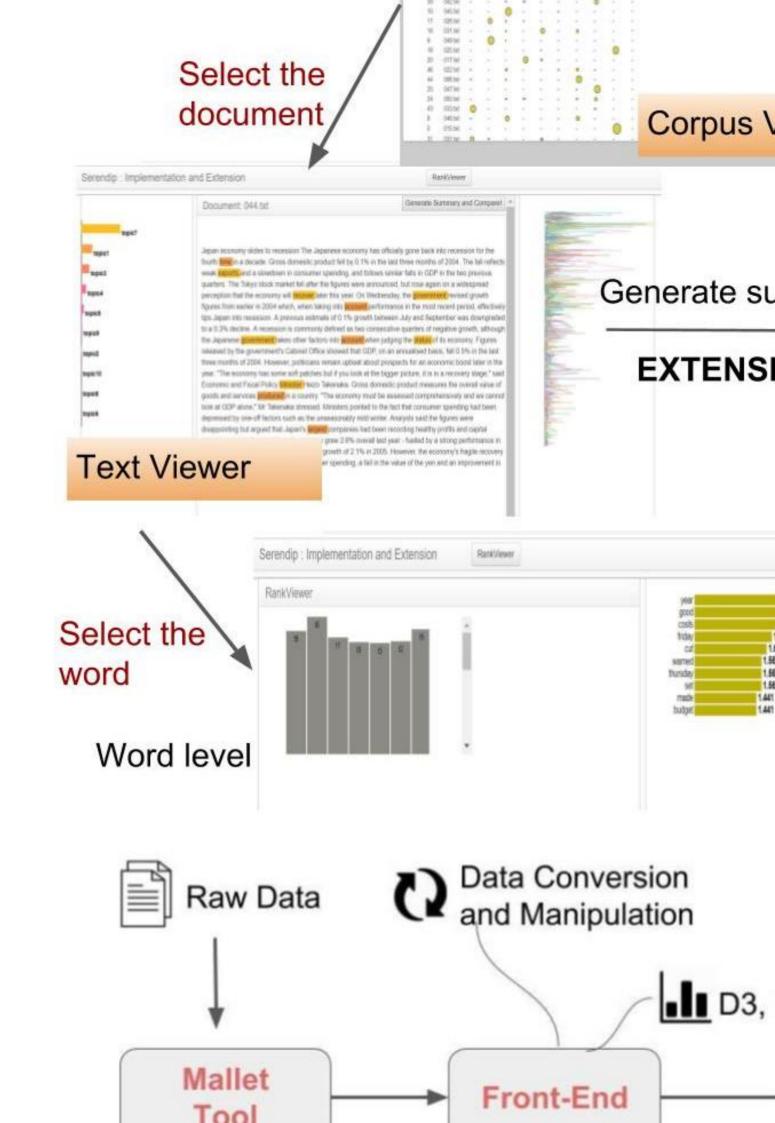
System Description

- Our model is a web a from Mallet tool, throu
- Corpus Viewer Doo documents VS topics by a bubble of corres and sorting.
- Text Viewer Text/P topics, document text each document, show word highlighting, and include selection, high
- DOCUMENT SUMM/ summary has been original document u
- Rank Viewer Word and the top 10 ranked selection and sorting.

tion of

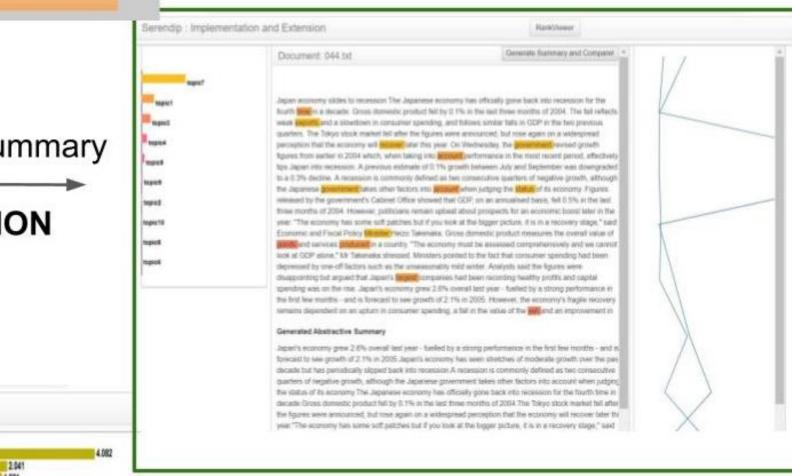


pplication showing topic-model data obtained ugh 3 views with task-dependent interactions. cument level view - a reorderable matrix of , weight of each topic in a document shown ponding radius. Interactions include selection aragraph level view - a three-part view of the , and word position in the document. For vs the top topics, the document itself with d line chart of the word position. Interactions hlighting and navigation to Rank Viewer. ARIZATION - Extension - Each document's generated and visually compared with the ising line chart to evaluate our model. level view - a simple view of all the topics d words in each topic. Interactions include



Disconnection of the control of the

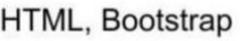
Comparing document and summary



Rank Viewer

Results:

- Datasets can be analy
- You can sort the docu
 Let's sort based on To
- Selecting this docume words in this docume
- The article is about de Cooper. Clicking on the

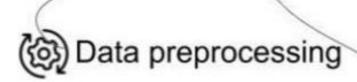


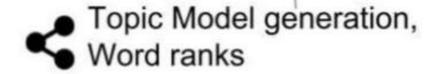


<u>Datasets:</u>

- BBC news comprising of articles dataset consists of 2225 documents from the BBC news website corresponding to stories in five topical areas from 2004-2005 and the five Class Labels are business, entertainment, politics, sport, tech.
- Movie scripts from the Internet Movie Script Database(IMDb), comprising of scripts from various genres and languages.
- Preprocessing and topic-model generation has been done in Mallet.

yzed starting from the Corpus Viewer.
Iments and topics on the Corpus Viewer.
Iments and topics on the Corpus Viewer.
Impic 2. 049.txt most belongs to Topic 2.
Int will take you to Text Viewer - all the int under Topic 2 are highlighted.
Internal of cars like BMW, Beetle and Minital of the BMW's will take you to Rank Viewer.





Amisha Thakkar (athakka2@asu.edu), A (hchikkal@asu.edu), Satya Srinija Kante Corpus Viewer Text Viewer Rank Viewer



- Rank Viewer shows s

 Topic 9 appears first,
 like 'year', 'prices', 'de
- Generate the summand also is about cars. The the line chart. Both line

Aravind Shanmugam (<mark>ashanmu5@asu.edu</mark>), l eti (<u>skantet1@asu.edu</u>) orted topics based on the word 'BMW'. and on clicking Topic 9, the words are all mand', just like the article talks about. Ty for 049.txt. You can see that the summary is can be compared with the original using es look similar for most topics.

