Document Snippet Generation

Troya Çağıl Köylü  21002018
Outline

- Introduction
- Phases
  - Preprocessing
  - Clustering
  - Labeling
  - Performance Measures
- Future Work
- Sample Run
- References
Introduction

Aim is to present retrieved documents in clustered & labeled manner to enable:

- Better topic understanding
- Ease of access
Phases

- Preprocessing
- Clustering
- Labeling
- Performance Measures
  - Clustering evaluation
  - Labeling evaluation
Preprocessing

- Database selection & preparation
- Document/Term matrix creation
  - Using Porter Stemmer
  - Binary or weighted?
year > year
make > make
making > make
coward > coward
cowardly > cowardly
note > note
noted > note
life > life
live > live
like > like
likely > like
unlikely > unlikely
propose > propose
proposes > proposes
cat > cat
cats > cat
Clustering

- Cover Coefficient-based Clustering
  - Number of clusters
  - Initial cluster centers
- k-means Clustering
  - Using term vectors to calculate cluster means
  - Stopping if no centroid changes
Labeling

- Using document titles to label clusters [2]
- Labeling via term weighting [1]
Performance Measures

- Clustering evaluation
  - F-measure (that uses recall & precision)
- Labeling evaluation
  - Human evaluation (ground truth)
  - sim-F measure [1]
Future Work

- Remaining phases
- Labeling
- Performance evaluation
- Modifications, changes & additions
Sample Run

Number of documents: 51
Number of terms: 2304
Number of clusters: 21
Number of iterations in k-means: 168
Execution time: ~2 seconds
Cluster 1: [world news in brief: brussels rioting]

Cluster 2: [world news in brief: cocaine ring broken, world news in brief: newspaper pays up]

Cluster 3: [world news in brief: malawi frees poet]

Cluster 4: [observer: nobbled]

Cluster 5: [world news in brief: population warning]

Cluster 6: [world news in brief: booby-trap murder]

Cluster 7: [london stock exchange: equity futures and options trading, london stock exchange: ft-se 2,500 lost in nervous trading, world stock markets (america): bond recovery fails to bring joy to equities, government bonds: gilts follow treasuries' fall in auction]

Cluster 8: [observer: free wheeling, the lex column: british airways, leading article: a compromise path to emu]

Cluster 9: [us economic hopes rise as output level stabilises, foreign exchanges: dollar and yen lose ground, commodities and agriculture: lme monitors copper market as squeeze tightens, commodities and agriculture: fuel cell use is predicted for platinum, international company news: fletcher challenge shares placed, uk consumer credit growth at six-year low, the lex column: uk credit]

Cluster 10: [international company news: sa brewing expands in us water heating, bush pledge on chemical weapons]

Cluster 11: [international company news: usf&g to bolster capital by dollars 300m stock issues, international capital markets: citicorp in asset-backed launch worth dollars 1.4bn, severn trent to pay pounds 212m for bet unit, drexel details settlement plan, letter: government cover for private export insurance]

Cluster 12: [international company news: labatt to reduce its food processing operations, commodities and agriculture: farmers' viewpoint - machinery, the lex column: severn trent places a bet on waste]

Cluster 13: [uk company news: geevor merger hits rocks over pre-conditions, jubilee of a jet that did what it was designed to do, commodities and agriculture: total in algerian gas field, international company news: temporary rescue for french, hurd opposes wider role for ec on labour issues, leading article: a charter for consumers]
Cluster 14: [international company news: contigas plans dm900m east german project, international company news: apple computer upgrades]

Cluster 15: [the lex column: stock lending]

Cluster 16: [brent walker shares drop ahead of loss, uk company news: business spread behind power corp's advance, tvs unveils plan to raise equity as it falls into red]

Cluster 17: [world news in brief: khmer rouge snub]

Cluster 18: [the lex column: uk markets]

Cluster 19: [world news in brief: soviet deadlock]

Cluster 20: [international company news: toyoda machinery sales surge, international company news: nordstrom boosts earnings, green profits]

Cluster 21: [international company news: two venezuelan airlines register losses, international company news: aerolneas argentinas sale enters final phase]
References
