Classification & Clustering of Dear Colleague Letters

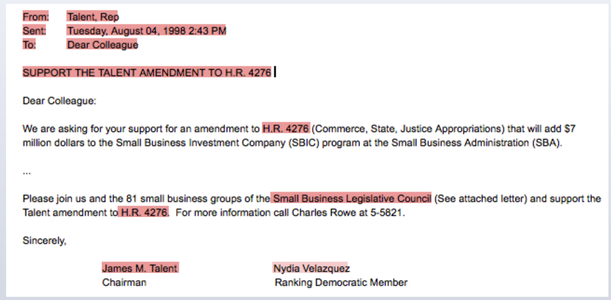
(with Professor D.Christenson)
Rushi Ganmukhi

Introduction

The Data: Dear Colleague Letters

Letters between House Representatives asking for support, opposition or cosponsorship on a bill. Often contain names of interest groups involved.

Want to find Relationships between representatives, organizations and bills



From: Talent, Rep
Sent: Tuesday, August 04, 1998 2:43 PM
To: Dear Colleague

SUPPORT THE TALENT AMENDMENT TO H.R. 4276

Dear Colleague:

We are asking for your support for an amendment to H.R. 4276 (Commerce, State, Justice Appropriations) that will add $7 million dollars to the Small Business Investment Company (SBIC) program at the Small Business Administration (SBA).

...

Please join us and the 81 small business groups of the Small Business Legislative Council (See attached letter) and support the Talent amendment to H.R. 4276. For more information call Charles Rowe at 5-5821.

Sincerely,

James M. Talent, Chairman
Nydia Velazquez, Ranking Democratic Member

Matching

Needed to match Representative, Organization and Bill Names present in the emails

Aderholt, Robert B
ADERHOLT, ROBERT B
Robert B Aderholt
ROBERT B ADERHOLT
Aderholt
ADERHOLT
Robert Aderholt
ROBERT ADERHOLT

'HR':'house_bill', 'H.R.':'house_bill',
'S.':'senate_bill', 'S':'senate_bill',
'S.Res.':'senate_joint_resolution',
'SRes':'senate_joint_resolution',

Clustering

AMERICAN MEDICAL ASSOCIATION
LIFE LEGAL DEFENSE FOUNDATION
EAGLE FORUM
AMERICAN PUBLIC HEALTH ASSOCIATION
LIFE ISSUES INSTITUTE
AMERICAN CONSERVATIVE UNION
PLANNED PARENTHOOD
TRADITIONAL VALUES COALITION
LIBERTY COUNSEL

NATIONAL WOMENS HEALTH NETWORK
GUTTMACHER INSTITUTE
CONSUMERS UNION
WOMEN
GOVERNMENT ACCOUNTABILITY PROJECT
FEMINIST MAJORITY FOUNDATION

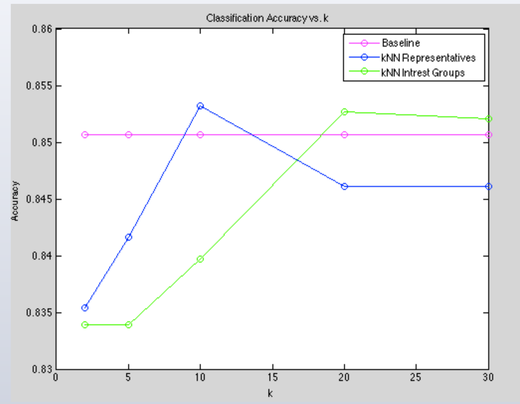
Used K-means to cluster Organizations and Representatives based on which bills linked to by the emails.
K = 2,5, 10, 100

Classification

K- Nearest Neighbors Classification

Wanted to predict how far a bill progresses (From govtrack.us)

Compared organizations and representatives present on each bill



Less-Successful Results

Sentiment Analysis:
Parser is trained on Movie Reviews

Not many sentiment or opinion words in the emails

LDA Topic Modeling:
Ran it on emails and bill titles
Sparsity of data in emails led to poor results.

#15: sense, expressing, house, act, representatives, united, national, states, trade, guard

#16: tax, act, code, amend, revenue, internal, 1986, credit, relief, provide

Frequency Analysis

SEC 78
SIERRA CLUB 87
INTERNATIONAL UNION 91
CONSUMERS UNION 97
NATIONAL EDUCATION ASSOCIATION 106
AMERICAN LEGION 113
AMERICAN MEDICAL ASSOCIATION 119
FEDERAL TRADE COMMISSION 121
ASSOCIATED PRESS 122

Scott, Bobby 1296
Schakowsky, Jan 1319
Conyers, John Jr 1325
Paul, Ron 1334
Long, Billy 1374
Filner, Bob 1384
McGovern, James P 1405
Grijalva, Raul M 1614
Jackson, Jesse Jr 1697
Frank, Barney 1861
John, Chris 4024