

# Linear Algebra in Document Summarization

Symposium in Honor of Dianne P. O'Leary  
SIAM Conference Applied Linear Algebra 2015

*John M. Conroy*  
IDA Center for Computing Sciences  
Bowie, MD  
USA



refugee crisis europe

Web

**News**

Videos

Images

Maps

More ▾

Search tools

About 6,340,000 results (0.35 seconds)



### **Refugee crisis: Slovenia calls in army to help patrol borders**

**The Guardian** - 6 hours ago

Slovenia on Tuesday called in the army to help it manage **refugees** seeking to reach northern **Europe** before winter, as the small EU state ...

### **Europe's Refugee Crisis: Slovenia to Deploy Its Army**

**The Atlantic** - Oct 20, 2015

### **Europe refugee crisis: Slovenia to use army to help control border flow**

**CBC.ca** - Oct 20, 2015

### **Slovenia calls army to border to handle asylum seeker surge**

Opinion - **ABC Online** - 20 hours ago

### **Europe refugee crisis: Slovenia may raise border fence against ...**

In-Depth - **Sydney Morning Herald** - Oct 20, 2015

### **The disintegration of Europe**

Opinion - **Times of Oman** - Oct 19, 2015



Sydney M...



The Atlantic



CBC.ca



Internation...



Deutsche ...



Irish Times

**Explore in depth** (1,783 more articles)



## A tool for querying, clustering, and summarizing documents.

[ [Main](#) ] [ [Advanced](#) ] [ [About](#) ] [ [Help](#) ]

Query: Documents:   

### Navigation:

- 1 Q:94 [MEDL12610357.S](#)  
     93 [MEDL11896121.S](#)  
     92 [MEDL12453327.S](#)
- C:1 [MEDL10824925.S:7](#)  
     2 [MEDL11376697.S:7](#)  
     3 [MEDL11861291.S:12](#)  
     4 [MEDL11325836.S:12](#)
- 2 Q:79 [MEDL9438854.S](#)  
     78 [MEDL11719439.S](#)  
     73 [MEDL11334388.S](#)
- C:1 [MEDL11850541.S:21](#)  
     2 [MEDL12094373.S:12](#)  
     3 [MEDL11850541.S:20](#)  
     4 [MEDL12094373.S:13](#)  
     5 [MEDL11850541.S:22](#)

### Multiple Document Summaries:

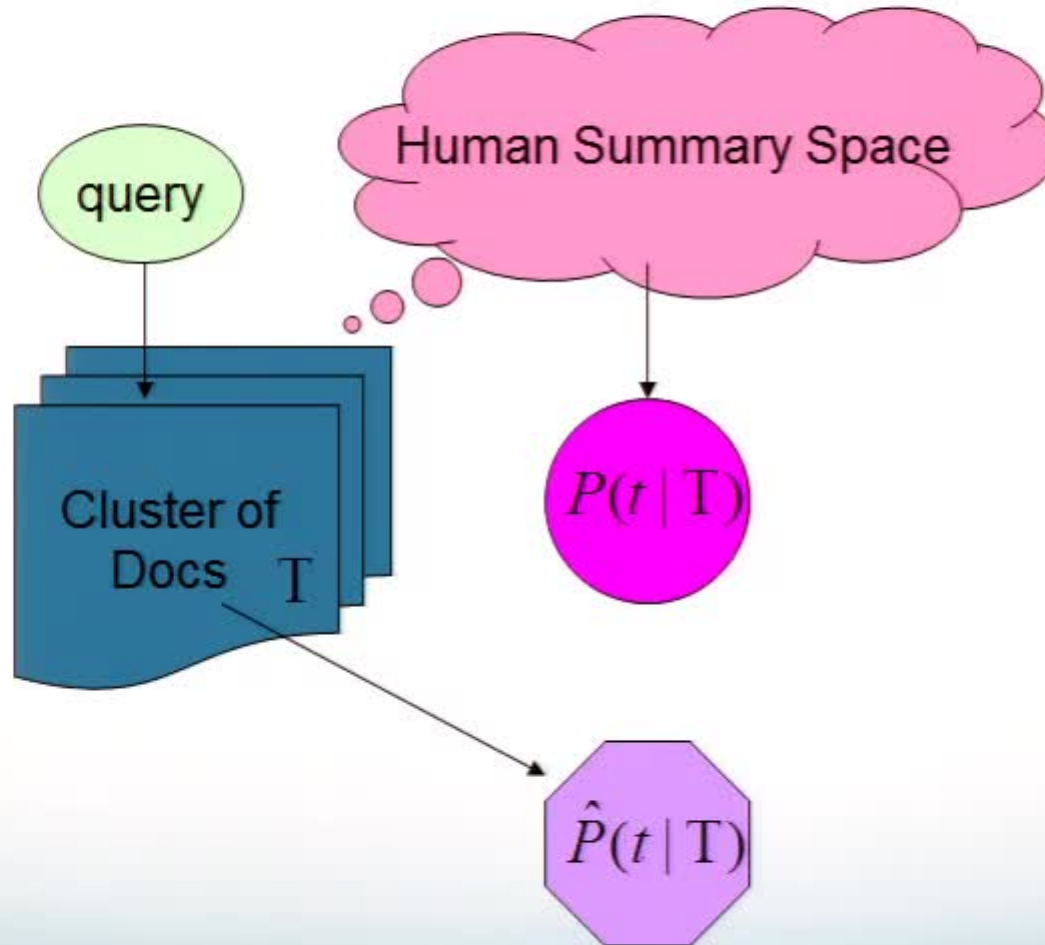
- 1 C-kit proto-oncogene product (KIT, CD117) is a tyrosine kinase growth factor receptor for stem cell factor. Loss-of-function mutations of the c-kit receptor tyrosine kinase (KIT) result in depletion of mast cells and interstitial cells of Cajal (ICCs). Mutations of c-KIT causing spontaneous activation of the KIT receptor kinase are associated with sporadic adult human mastocytosis (SAHM) and with human gastrointestinal stromal tumors. Six indolinone tyrosine kinase inhibitors were characterized for their ability to inhibit Kit kinase and for their effects on the growth of small cell lung cancer (SCLC) cell lines.  
**Mean Score: 85 Number of documents: 19**
- 2 There were c-kit mutations in 66.6% of benign GISTs (14/21), 83.3% of the malignant (5/6), and 40% of the cases of intermediate malignancy (2/5). In vitro studies suggest that GISTs with regulatory-region KIT mutations are more likely to respond to STI-571 than are GISTs with enzymatic-region mutations. In both Exons 9 and 13, single mutations could be identified, whereas no mutations were found in Exon 17. A minority of GISTs lack demonstrable KIT mutations, but KIT is nonetheless strongly activated. A low frequency of mutations in benign GISTs, as reported previously by other researchers, could not be observed in our panel.  
**Mean Score: 62 Number of documents: 20**

# Summarization Task

- Given a cluster of documents, create a **fluent** synopsis of specified length using information in the document cluster.
- Model Problem:
  - English newswire documents.
- **Other Applications:**
  - technical documents, Multi-lingual, email, speech transcripts.
- Can also be used to improve information retrieval.

T. Sakai and K. Spärk Jones, "Generic summaries for indexing in information retrieval", *SIGIR 01, Proceedings of 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001.

# A Model of Human Abstracts



# An Approximation of $P(t|\tau)$

- We approximate  $P(t|\tau)$  by

$$P_{qs\rho}(t | \tau) = \alpha_q q(t) + \alpha_s s(t) + \alpha_\rho \rho(t)$$

$$s(t)[q(t)] = \begin{cases} 1 & \text{if } t \text{ is a signature [query] term} \\ 0 & \text{if } t \text{ is not a signature [query] term} \end{cases}$$

$\rho(t | \tau) =$  probability  $t$  occurs in a sentence considered for selection.

- The score of a sentence is the sum of  $P(t)$  taken over its terms divided by its length.
- Conroy, O' Leary, Schlesinger ACL/COLING 2006.

# An Approximation of $P(t|\tau)$

- We approximate  $P(t|\tau)$  by

$$P_{qs\rho}(t | \tau) = \alpha_q q(t) + \alpha_s s(t) + \alpha_\rho \rho(t)$$

$$s(t)[q(t)] = \begin{cases} 1 & \text{if } t \text{ is a signature [query] term} \\ 0 & \text{if } t \text{ is not a signature [query] term} \end{cases}$$

$\rho(t | \tau) =$  probability  $t$  occurs in a sentence considered for selection.

- The score of a sentence is the sum of  $P(t)$  taken over its terms divided by its length.
- Conroy, O' Leary, Schlesinger ACL/COLING 2006.

# Two Approaches

- Pivoted  $QR$ .
  - Conroy & O'Leary SIGIR 2001
- A “L1-non-negative” “ $QR$ ,”  $NR$  factorization
  - Non-negative factorization  $A \approx NR$ .
  - Generally “pre-condition” matrix by using a truncated SVD to remove noise.
  - Conroy, O' Leary, Schlesinger ACL/COLING 2006.



# Pivoted QR

Use scores to select candidate sentences ( $\sim 6w$ )

- Terms as sentence features
  - Terms:  $\{t_1, \dots, t_m\} \in \mathbf{R}^m$
  - Sentences:  $\{s_1, \dots, s_n\} \in \mathbf{R}^n$
  - Scaling:  $\|\mathbf{a}\|_2 = \text{score for sentence.}$
- Perform QR with column pivoting, implicit elimination to preserve sparsity.
- Stop criteria: chosen sentences (columns) have  $\sim w$  words.

|          | $s_1$    | $\dots$  | $s_n$    |
|----------|----------|----------|----------|
| $t_1$    | $a_{11}$ | $\dots$  | $a_{1n}$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $t_m$    | $a_{m1}$ | $\dots$  | $a_{mn}$ |

# QR Inspired Non-negative Matrix Factorization

Use scores to select candidate sentences ( $6w$  words).

- Terms as sentence features
  - Terms:  $\{t_1, \dots, t_m\} \in \mathbf{R}^m$
  - Sentences:  $\{s_1, \dots, s_n\} \in \mathbf{R}^n$
  - Scaling:  $\|\mathbf{a}\|_1$  = score for sentence.
- Maintain non-negativity but we lose orthogonality:  $\mathbf{A} \approx \mathbf{NR}$ .
- Choose column with maximum **1-norm** ( $\mathbf{a}_j$ )
- Subtract components along  $\mathbf{a}_j$  from remaining columns, **but set negatives to 0**.
- Stop criteria: chosen sentences (columns) have  $\sim w$  words

|          | $s_1$    | $\dots$  | $s_n$    |
|----------|----------|----------|----------|
| $t_1$    | $a_{11}$ | $\dots$  | $a_{1n}$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $t_m$    | $a_{m1}$ | $\dots$  | $a_{mn}$ |

# Evaluation Methods

- **Manual Evaluation**
  - Overall Responsiveness (grade for summaries 1-5 scale)
  - Pyramid (a content metric)
- **Automatic Evaluation**
  - ROUGE-1 (unigrams)
  - **ROUGE-2 (bigrams)**
  - ROUGE-3 (trigrams)
  - ROUGE-4 (tetra-grams)

Lin, Chin-Yew and E.H. Hovy 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003), Edmonton, Canada, May 27 - June 1, 2003.

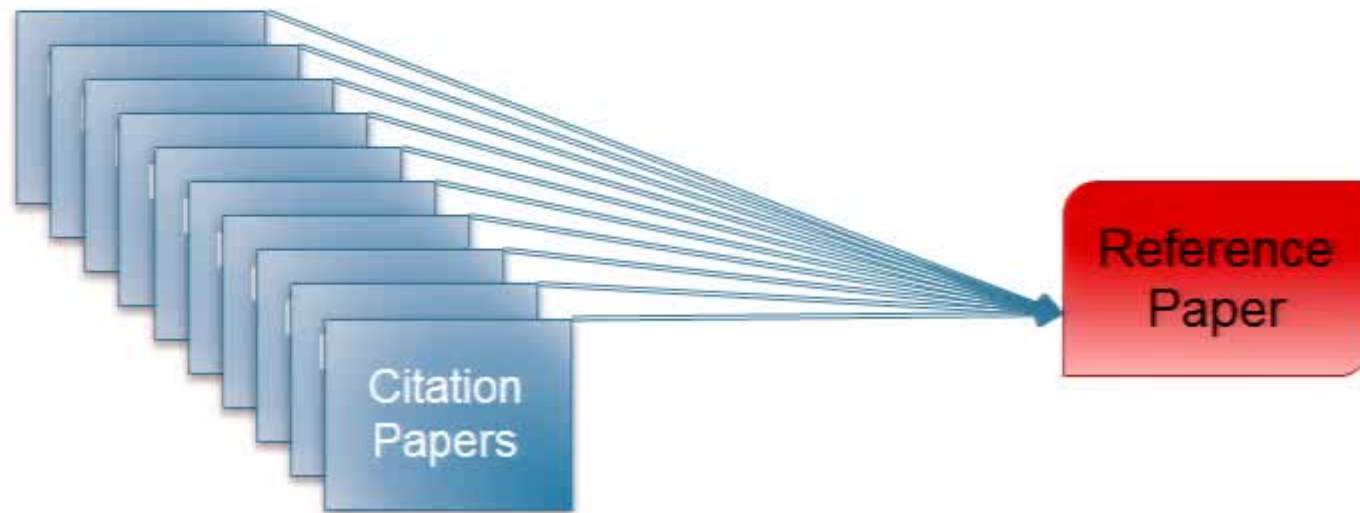
# Multilingual Summary Results (100 words)

| Selection Method           | ROUGE-2 | Confidence Interval |
|----------------------------|---------|---------------------|
| Pivoted <i>QR</i>          | 0.156   | (0.143,0.169)       |
| <i>NR</i><br>Factorization | 0.179   | (0.166,0.192)       |

# Multilingual Summary Results (100 words)

| Selection Method           | ROUGE-2 | Confidence Interval |
|----------------------------|---------|---------------------|
| Pivoted <i>QR</i>          | 0.156   | (0.143,0.169)       |
| <i>NR</i><br>Factorization | 0.179   | (0.166,0.192)       |

# From News to Science: Summarize a Multiply Cited Paper



Text Analysis Conference (TAC) 2014 Pilot Task:  
Biomedical Summarization  
20 papers, each with 10 papers citing it.

Vahed Qazvinian and Dragomir R. Radev. COLING 2008  
used citing sentences (citances)

At TAC 2014 human were allowed to use citances and  
reference paper.

# Data Processing and Segmentation

- Reference paper is processed,
  - Individual sections of the paper are isolated and extracted, and the document is then sentence split.
  - Sections from reference paper keep:
    - Abstract
    - Results
    - Other parts excluding bibliography.
- For each citation paper one or more citing sentences (*citances*) are extracted.

# Term Weighting to Covering

- Resulting term weights, computed by NNMF or LM are correlated with the probability that a human will include the term in a summary.
- *Then, OCCAMS*, an Optimal Combinatorial Covering Algorithm for Multi-document Summarization is then used to select a subset of the sentences, with at most 250 words, with near optimal covering.

Sashka T. Davis, John M. Conroy, and Judith D. Schlesinger. *ICDM Workshops 2012*.

John M. Conroy, Sashka T. Davis, Jeff Kubina, Yi-Kai Liu, Dianne P. O'Leary, and Judith D. Schlesinger. ACL 2013. In *MultiLing Workshop*



# OCCAMS\_V5 Sentence Selection

- An optimal combinatorial covering algorithm for multi-document summarization.

Algorithm **OCCAMS\_V** ( $T, \mathcal{D}, \mathcal{W}, c, L$ )

1.  $K_1 = \text{Greedy\_BMC}(T, \mathcal{D}, \mathcal{W}, c, L)$
2.  $K_2 = S_{max} \cup \text{Greedy\_BMC}(T', \mathcal{D}', \mathcal{W}, c', L')$ ,  
where  $S_{max} = \text{argmax}_{\{S_i \in \mathcal{D}\}} \left\{ \sum_{t_j \in S_i} w(t_j) \right\}$   
and  $T', \mathcal{D}', \mathcal{W}, c', L'$  represent quantities updated  
by deleting sentence  $S_{max}$  from the collection.
3.  $K_3 = \text{KS}(\text{Greedy\_BMC}(T, \mathcal{D}, \mathcal{W}, c, 5L), L)$ ;
4.  $K_4 = \text{KS}(K'_4, L)$ , where  
 $K'_4 = S_{max} \cup \text{Greedy\_BMC}(T', \mathcal{D}', \mathcal{W}, c', 5L')$ ;
5.  $K = \text{argmax}_{k=1,2,3,4} \left\{ \sum_{T(K_i)} w(t_i) \right\}$   
where  $T(K_i)$  is the set of terms covered by  $K_i$ .

# Results for NNMF Weights

| System     | R1           | R2           | R3           | R4           |
|------------|--------------|--------------|--------------|--------------|
| TF         | 0.511        | 0.166        | 0.065        | 0.030        |
| NNMF_2     | 0.509        | 0.172        | 0.073        | <b>0.036</b> |
| NNMF_4     | 0.504        | 0.171        | <b>0.074</b> | 0.036        |
| NNMF_35    | <b>0.518</b> | <b>0.176</b> | 0.070        | 0.033        |
| Avg Human  | <b>0.528</b> | <b>0.179</b> | <b>0.075</b> | <b>0.036</b> |
| Best Human | 0.572        | 0.219        | 0.110        | 0.071        |

Table 1: Vector Space Model based on TF and NNMF

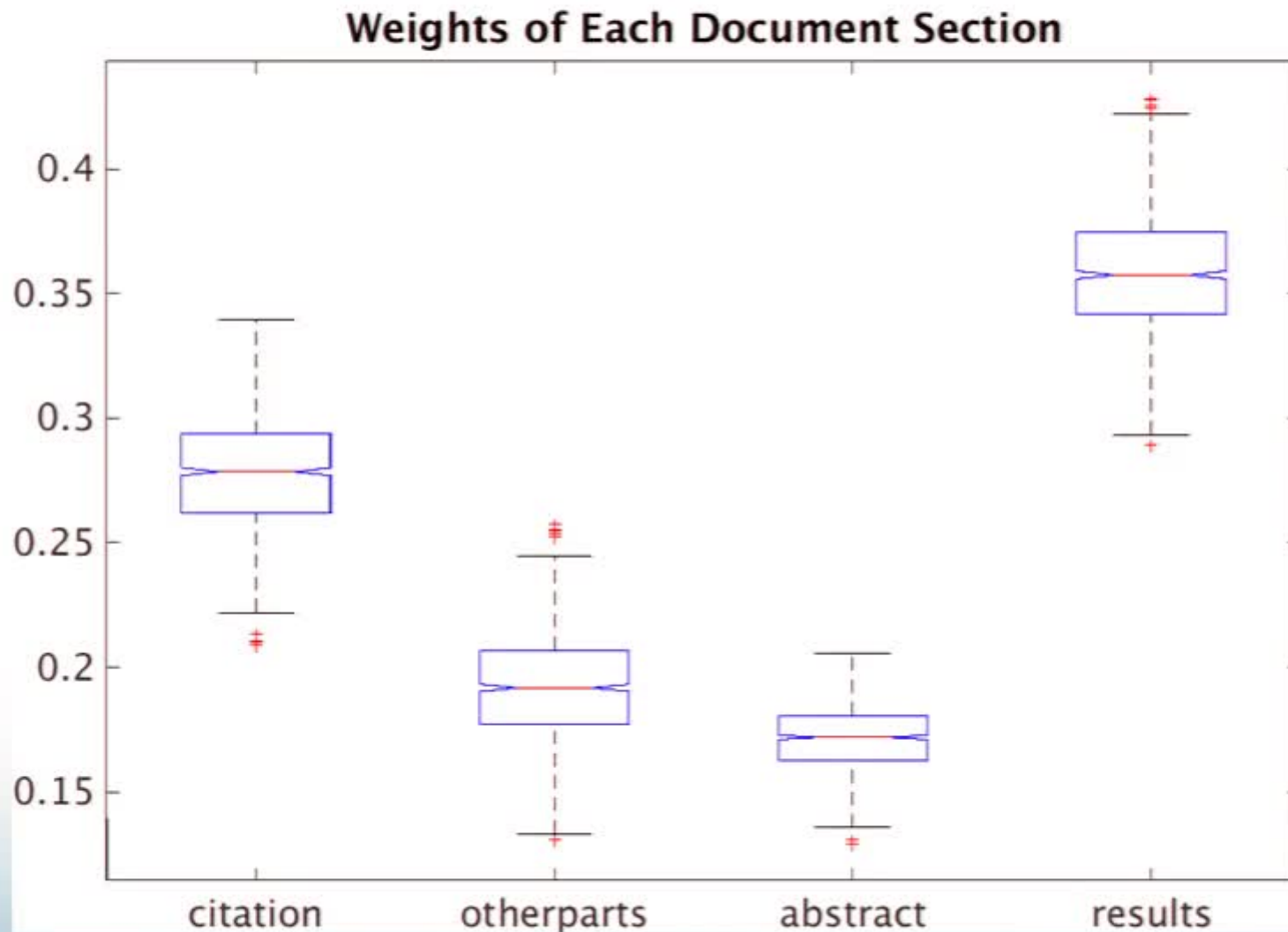
Rankel et al, ACL 2013, R3 and R4 give more discriminant power between systems than R1 or R2.

# Results for Language Models

| System         | R1    | R2    | R3    | R4    |
|----------------|-------|-------|-------|-------|
| $LM_1$         | 0.511 | 0.169 | 0.067 | 0.031 |
| $LM_4^{equal}$ | 0.559 | 0.210 | 0.095 | 0.052 |
| $LM_4^{opt}$   | 0.562 | 0.216 | 0.100 | 0.055 |
| Avg Human      | 0.528 | 0.179 | 0.075 | 0.036 |
| Best Human     | 0.572 | 0.219 | 0.110 | 0.071 |

Table 2: ROUGE Results for Three Language Models and a Comparison to Human Performance

# LM Stability of Weights



JSD optimization using 1000 random 10-subsets of the 20 documents sets.