How Document Properties Affect Document Relatedness Measures

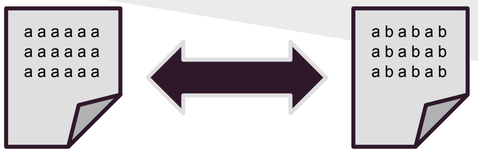
Jessica Perrie, Aminul Islam, Evangelos Milios
Dalhousie University, Faculty of Computer Science

Presented by **Diana Inkpen**
University of Ottawa,
School of Electrical Engineering and Computer Science

Introduction

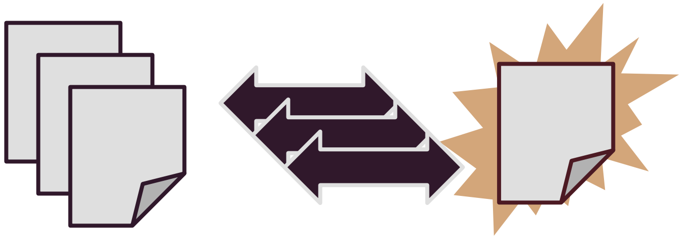
Document Relatedness

- Measurement of similarity…
- Between documents



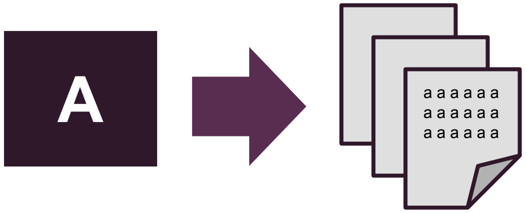
Applications: Document…

- Retrieval
- Clustering
- Classification
- Summarization



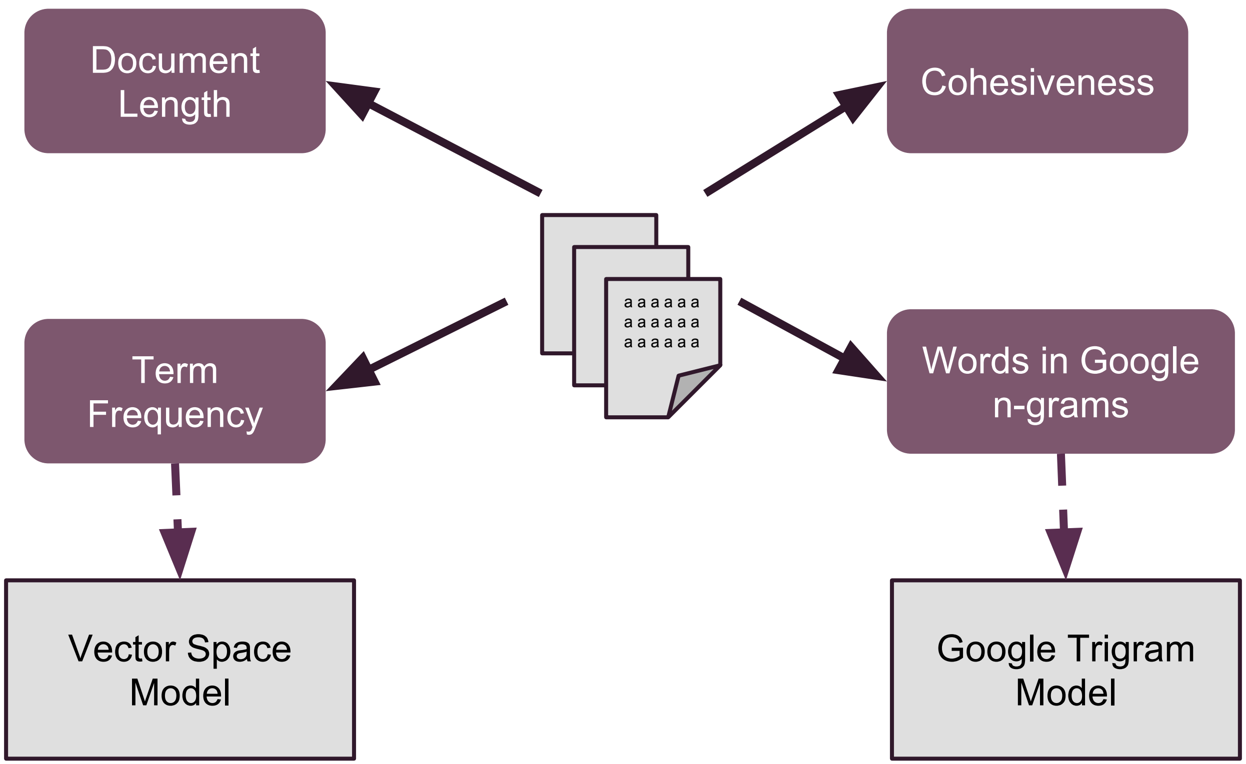
Unsupervised approaches

- Lack of training set requirement
- Performance depending on corpus



Motivation

Performance of a document relatedness approach depends on document properties -- found in the dataset being tested.



Contributions

General contributions:

- Presentation of different evaluations of document relatedness approaches on <u>different datasets</u>

selected based on their properties

- <u>Evaluations</u> based on intrinsic similarity of classes

*k*NN-classification

From experimental results:

- Evidence that different properties of documents yield better results in <u>different approaches</u>

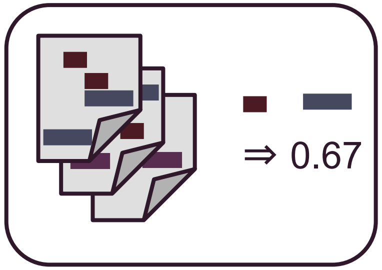
Vector Space Model & Google Trigram Model

[1] Islam, A., Milios, E., Vlado K.: Comparing Word Relatedness Measures Based on Google n-grams. In: 24th International Conference on Computational Linguistics, Proceedings of the Conference. COLING (Posters) '12.

Related Work

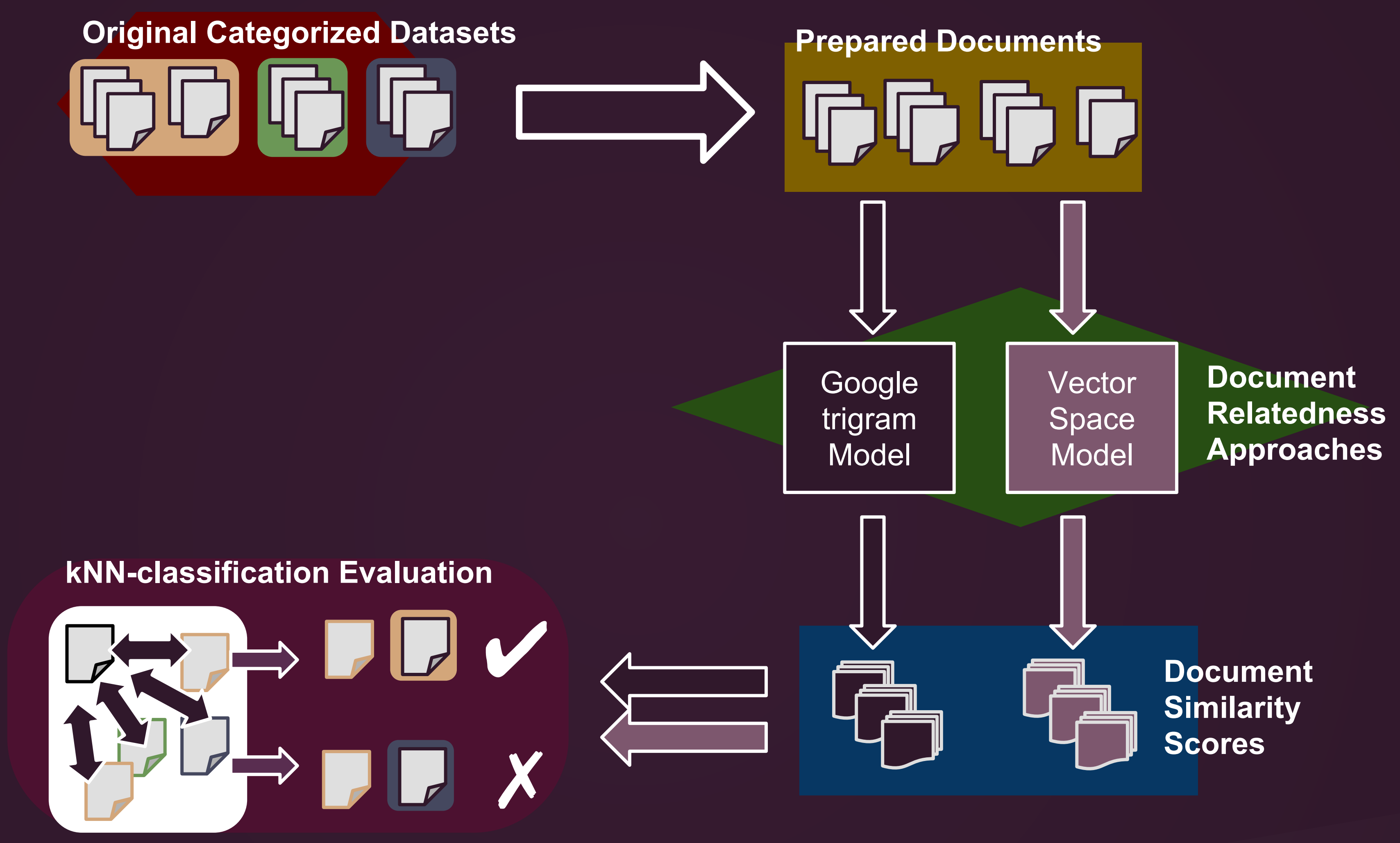
Unsupervised Corpus-based Approaches:

- Document Similarity Approaches
  - May use word similarity in back-end
- Word Similarity Approaches
  - Co-occurrence statistics
  - Corpus: web, dataset

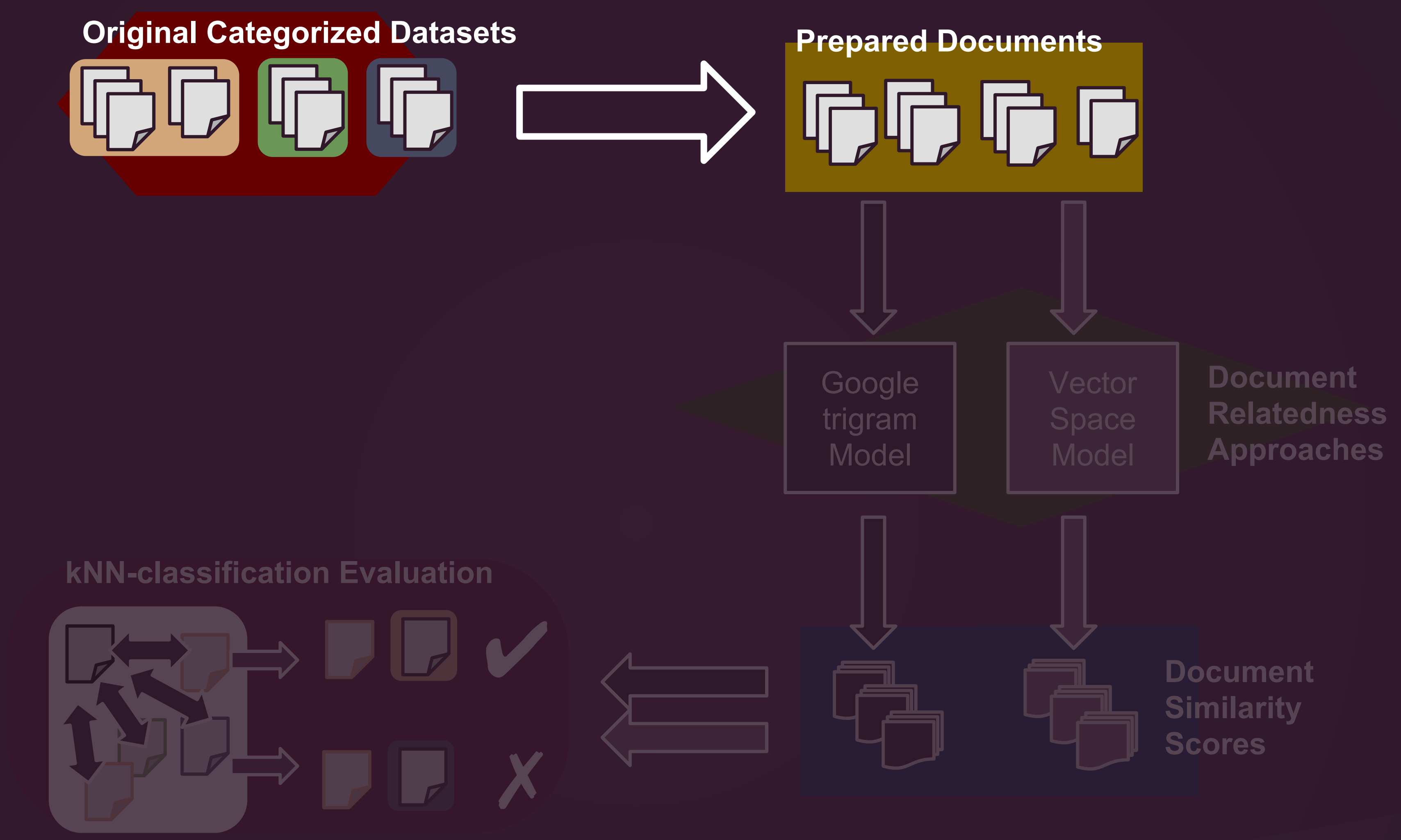


Comparisons of Different Approaches:

- Comparison of unsupervised corpus-based measures
  - Over human ratings, synonym tests
  - Measured: correlation, #correct synonyms
  - Text similarity with diff. word-relatedness approach



Methodology

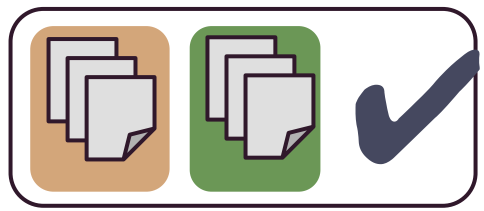

Original Categorized Datasets
Prepared Documents
Google trigram Model
Vector Space Model
Document Relatedness Approaches
kNN-classification Evaluation
✔
✗
Document Similarity Scores

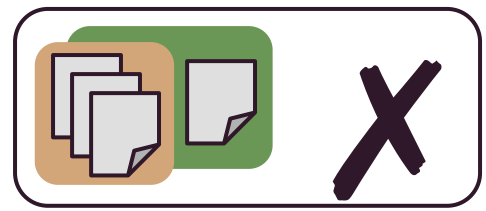
Datasets

Datasets

Document Relatedness defined:

- Document
- 1 category





Document Cleaning:

- Transform to lowercase, [^a-z] removal

Profits in poultry. Useful and ornamental breeds, and their profitable management.

profits in poultry useful and ornamental breeds and their profitable management

- Remove of 500+ English stop words

profits in poultry useful and ornamental breeds and their profitable management

profits poultry ornamental breeds profitable management

[2] https://c3.nasa.gov/dashlink/resources/138/

Datasets: ASRS

Aviation Safety Reporting System (ASRS)

- From SIAM 2007 Text Mining competition
- 22 categories total, mult-category assignment
- Over 4000 different words were concatenated together

⇒ **Selection:** 399 documents
⇒ **Document:** A single ASRS report
⇒ **Category:** Report's assigned category (4)
⇒ **Example:**

receive predepartureclearance AND setup WRONG depart ON flightmanagementsystem.NO aircraft conflict AND airtrafficcontrol indicate NO problem.

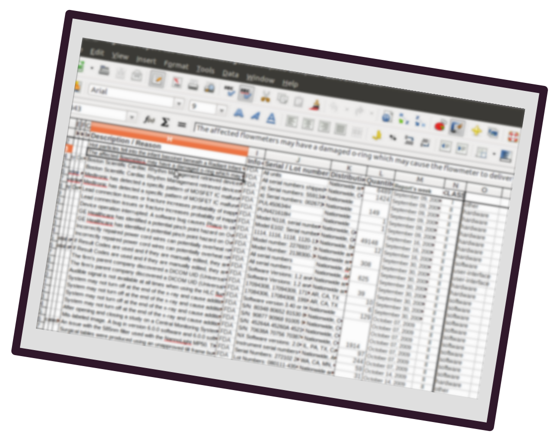
TEXT MINING 2007

Datasets: Med

Vigilance Report List (Med)

- Description for issues with medical equipment
- Provides reason for malfunction & subsequent categorization

⇒ **Selection:** 659 rows (367 unique)
⇒ **Document:** Description / Reason
⇒ **Category:** Categorization (2)
⇒ **Example:**



> Incorrect value calculations by the device may result in inaccurate aortic stenosis estimates

[3] http://www.biodiversitylibrary.org/
(OCR Description: http://biodivlib.wikispaces.com/BHL+and+Gaming)
(Example from: http://biodiversitylibrary.org/page/358022)

Datasets: BHL

Biodiversity Heritage Library (BHL)

- Biodiversity literature: pages, titles, subjects, authors
- Book text is Optical Character Recognition generated:

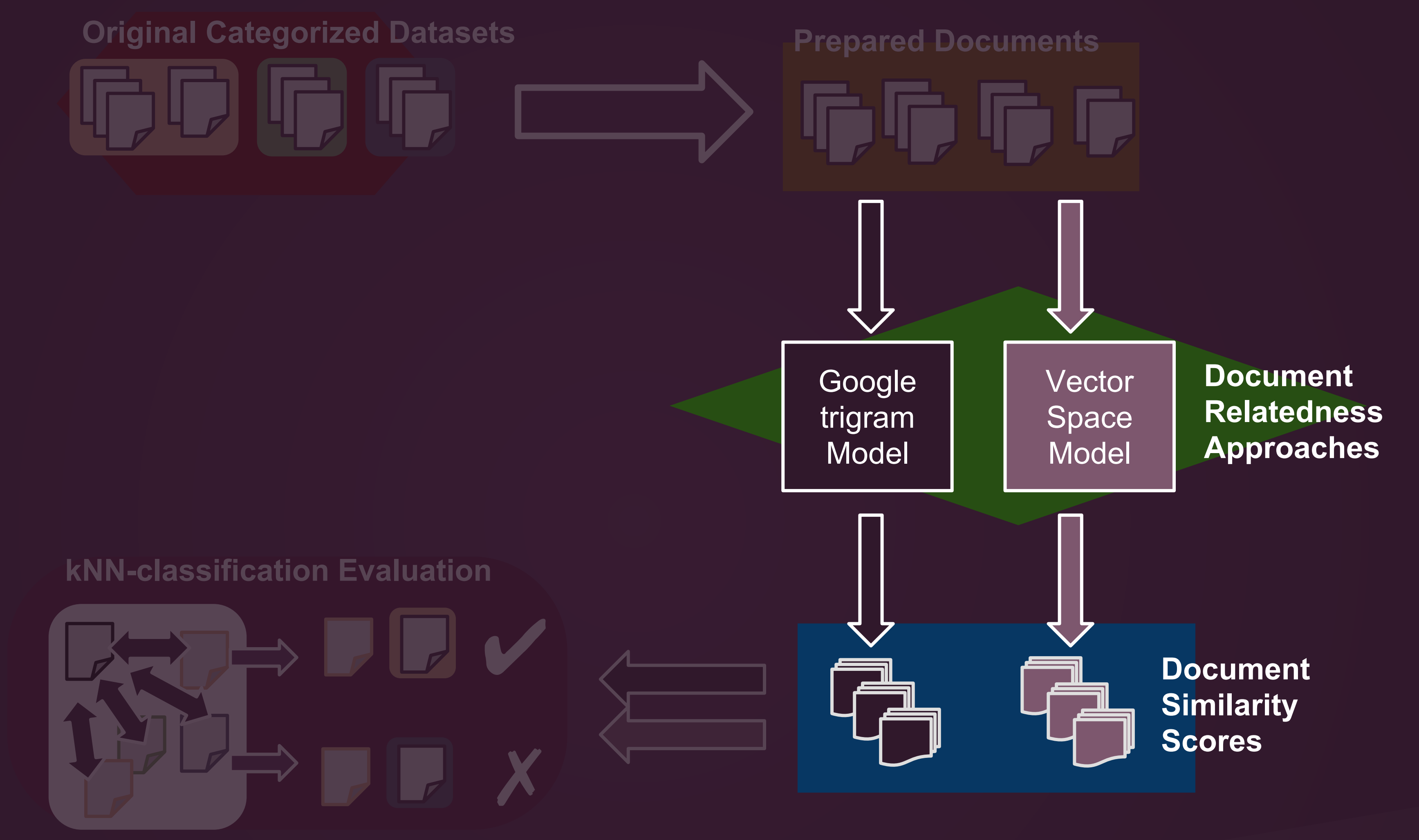
	Titles	**Intro**
⇒ **Selections:**	1152,	338
⇒ **Document:**	title,	contents' table, intro, preface
⇒ **Category:**	(4) subjects,	(5) subjects

⇒ **Examples:**

TABLE 6. SPECTROSCOPIC STANDARD OF CAROTIN AND XANTHOPHYLLIS. (FROM THE CARROT.) It will be noticed that the relative position of the bands of car- otin and xanthophylls is more […]

The vineyards of the world.

BHL
Biodiversity Heritage Library



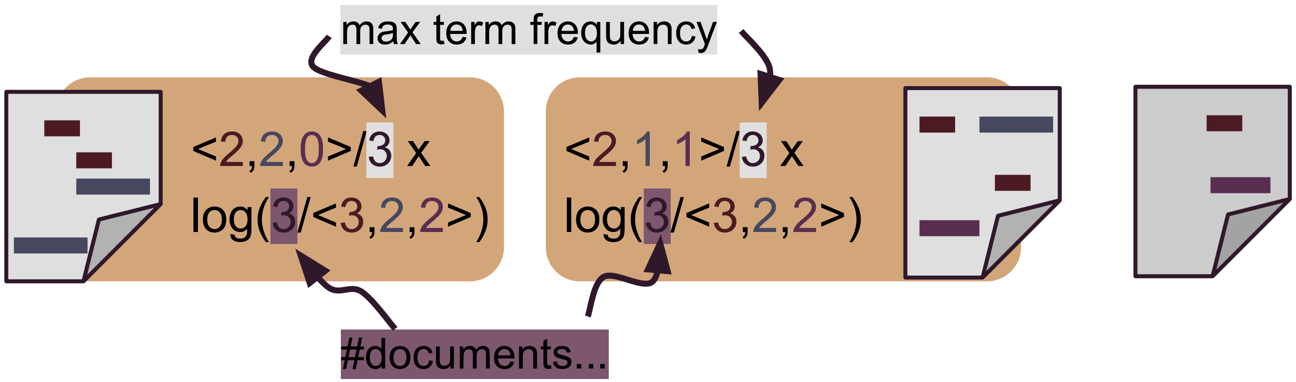
Unsupervised Corpus-based Approaches to Document Relatedness

[4] http://www.site.uottawa.ca/~diana/csi4107/cosine_tf_idf_example.pdf

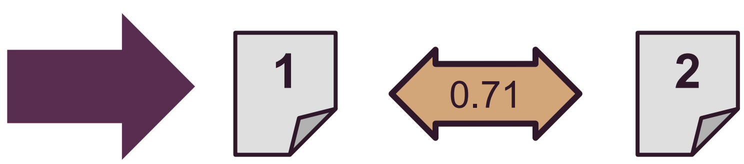
Approaches: VSM

Vector Space Model (VSM):

- Each document: vector with weights for each word
- Each weight: term-freq inverse doc-freq (TFIDF):



- Document relatedness: calculate cosine similarity

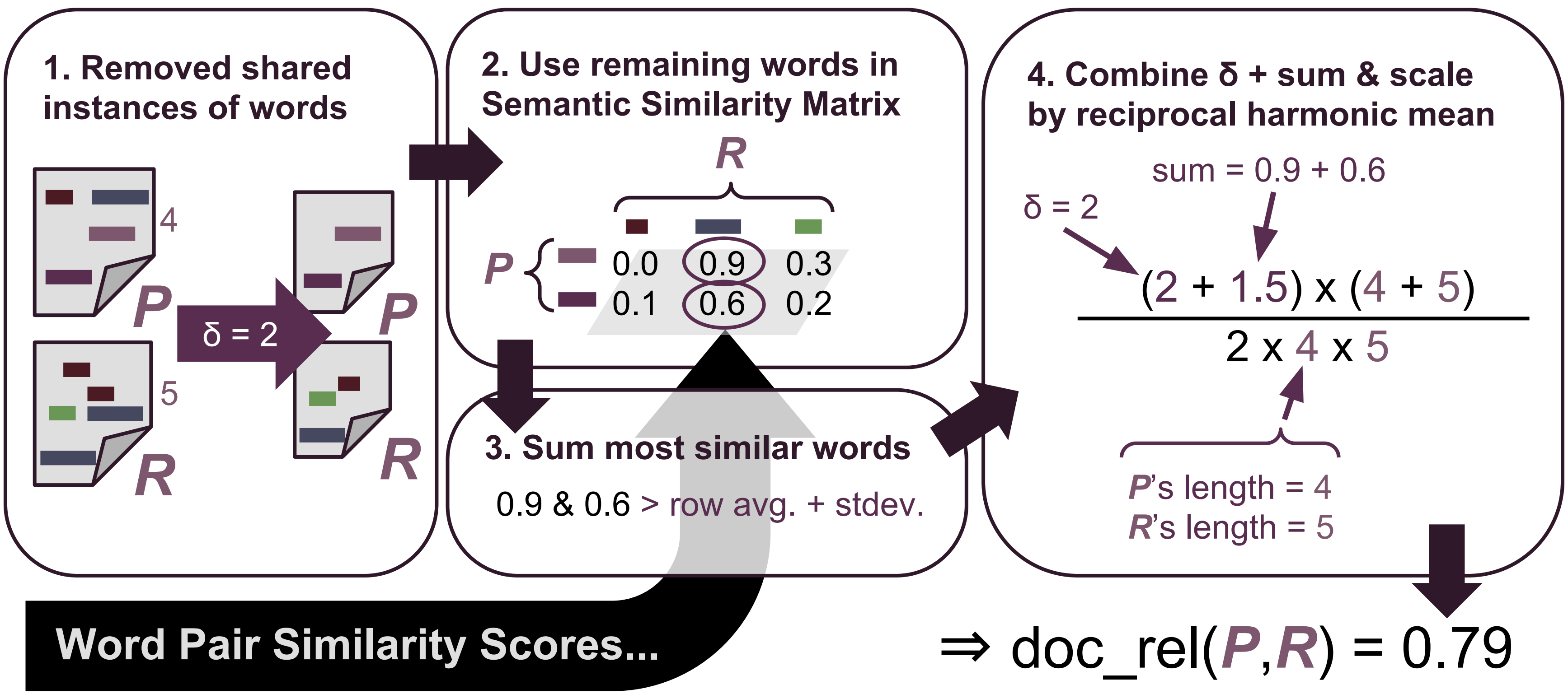


[7] Islam, A., Milios, E., Vlado K.: Text Similarity using Google Tri-grams. In: Advances in Artificial Intelligence; Lecture Notes in Computer Science. '12.

Approaches: GTM

Google Trigram Model (GTM):

- Document Relatedness: Use the shorter document's words & the longer document's most similar words

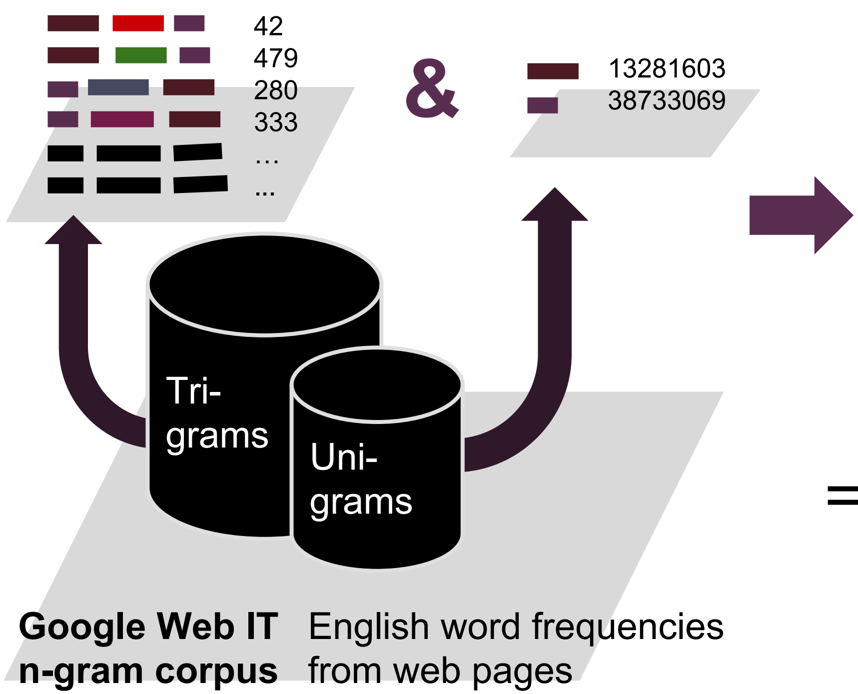


[7] Islam, A., Milios, E., Vlado K.: Text Similarity using Google Tri-grams. In: Advances in Artificial Intelligence; Lecture Notes in Computer Science. '12.

Approaches: GTM

GTM- Word Similarity:

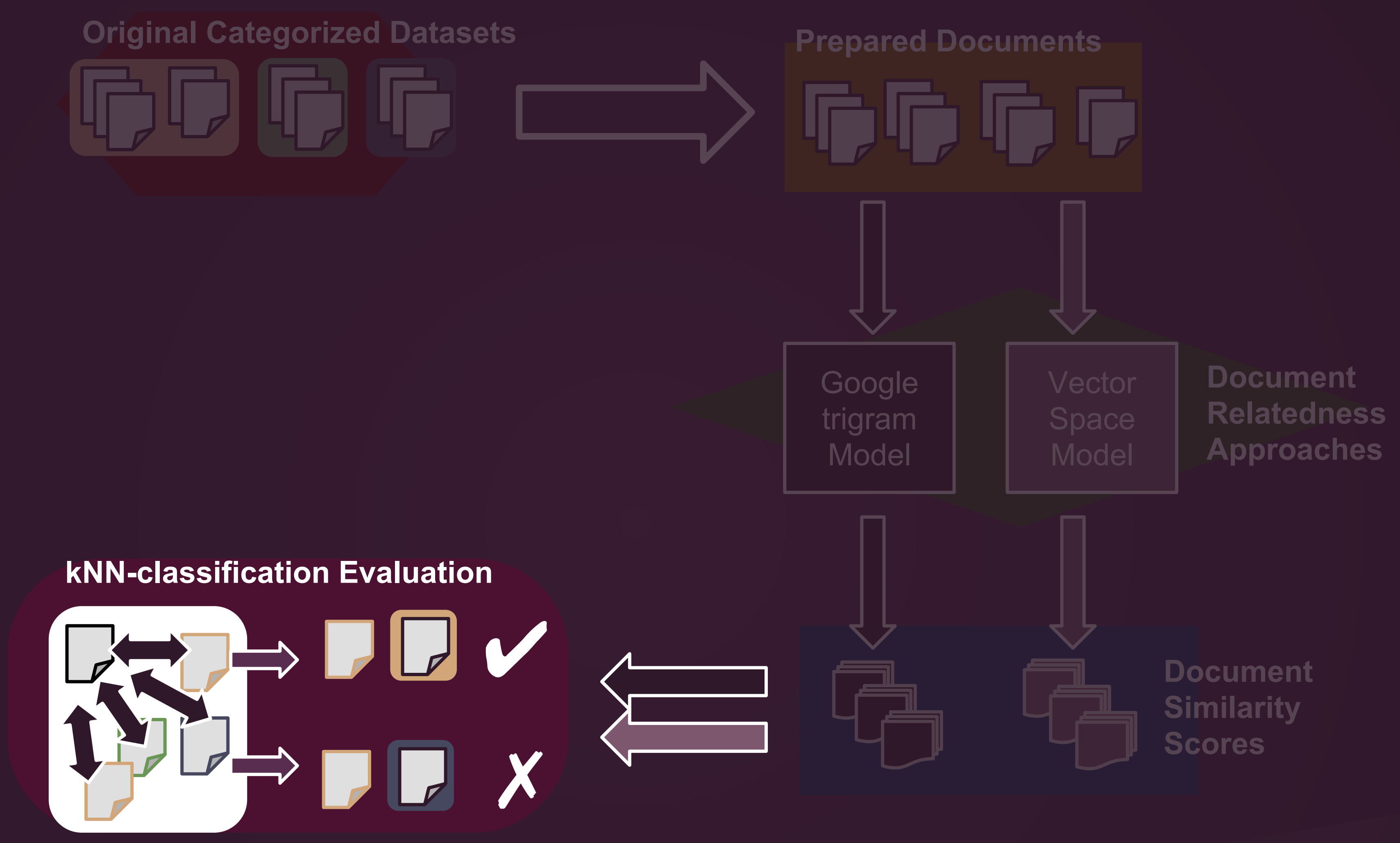
- Using Google trigrams & unigrams to calculate individual word similarity for word pairs



1. Find all trigrams that begin & end with pair
2. Normalize mean frequency

⇒ word_sim(▬ , ▬) = 0.52

Use of the GTM is available: http://ares.research.cs.dal.ca/gtm/



Evaluation: kNN-classification

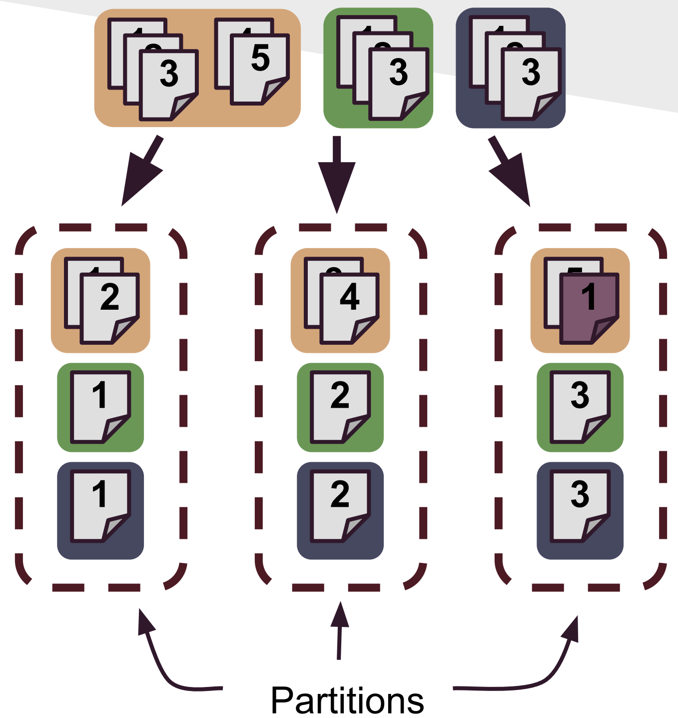
Evaluation: Setup

Representative Division:

Testing requires **30** different rand. generated partitioning

- 10-fold cross-validation
- Ea. partition = representative sample, some overlap

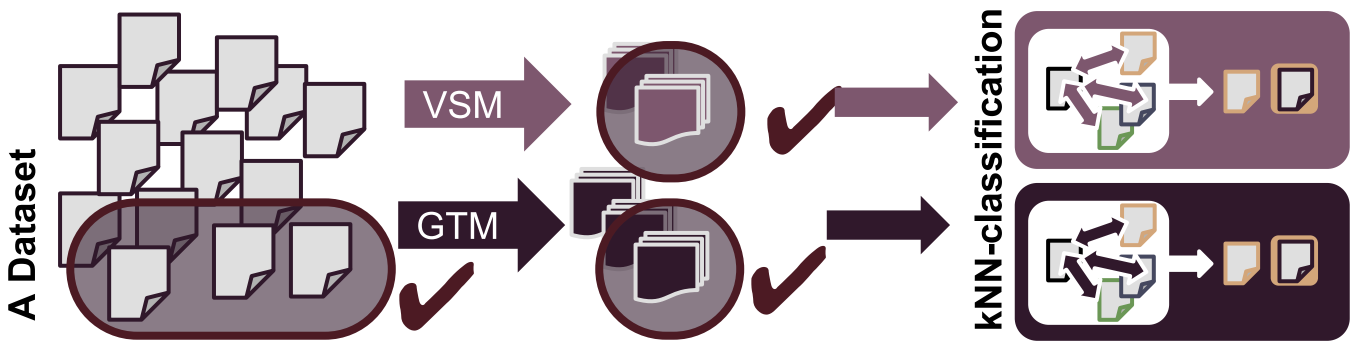
Ignored if neighbours



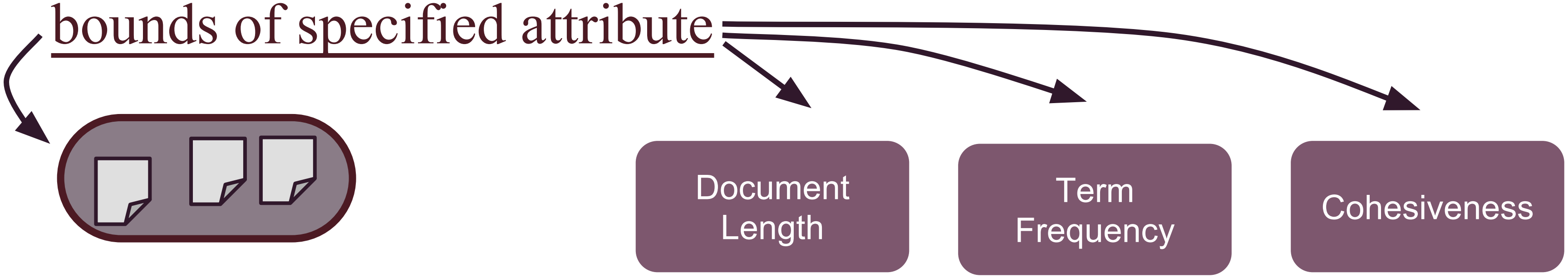
- Consider different k from $[1, \sqrt{\#\ \text{testing set documents}}\,]$
- Select k where mean accuracy is highest $\rightarrow$ accuracy

kNN-Classification

Dividing Dataset Similarity Scores



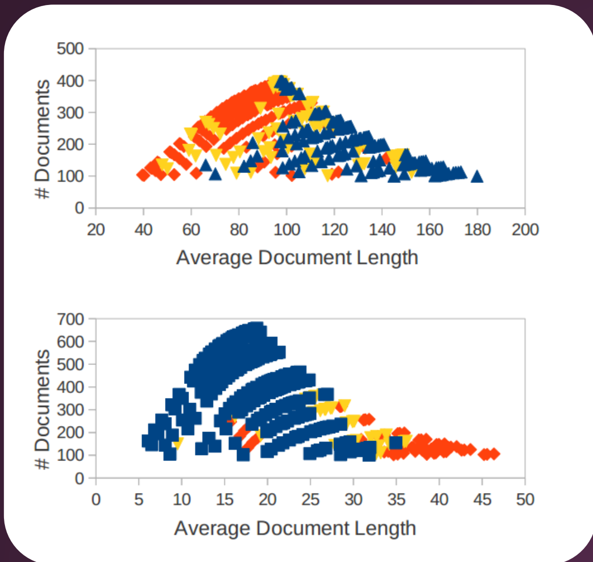
- Scores generated from documents within ea. dataset
- Run both evaluations on <u>only the documents that fit between bounds of specified attribute</u>



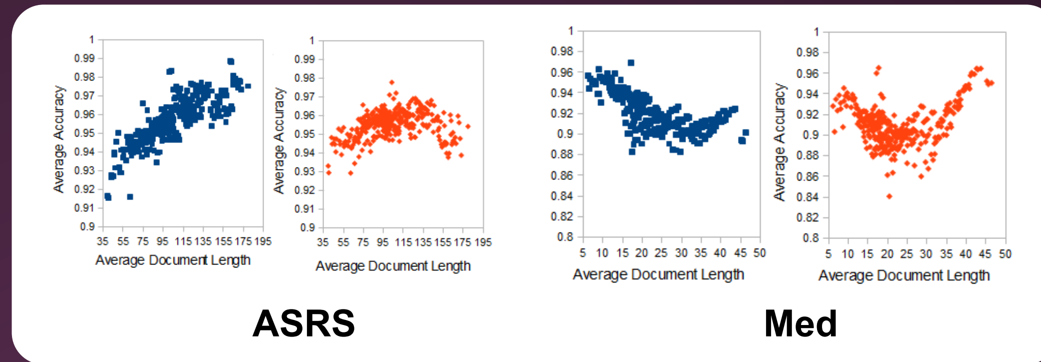
Results : GTM > VSM & VSM > GTM

ASRS

Med



Results : GTM & VSM



➢ Passed threshold → accuracies were higher for the other approach

ASRS

- VSM: shorter / longer documents: too few / too many words
- GTM: accuracy has a moderately a strong linear relationship

Med

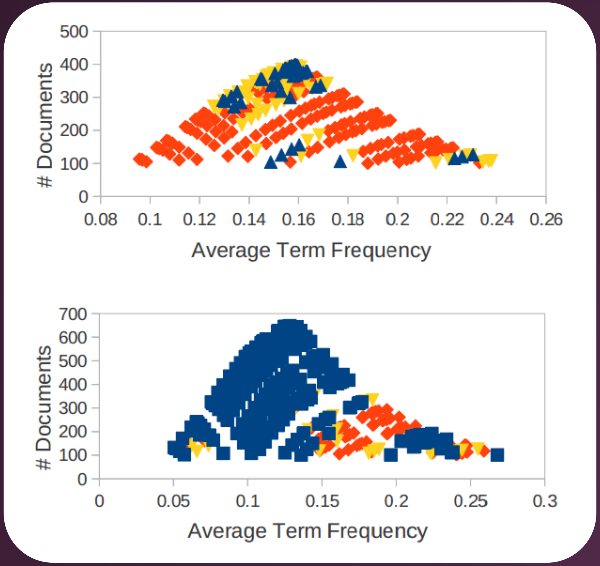
- At higher & lower bounds, similar documents helped accuracy

Comparison of Trigram & VSM Approaches: based on document length

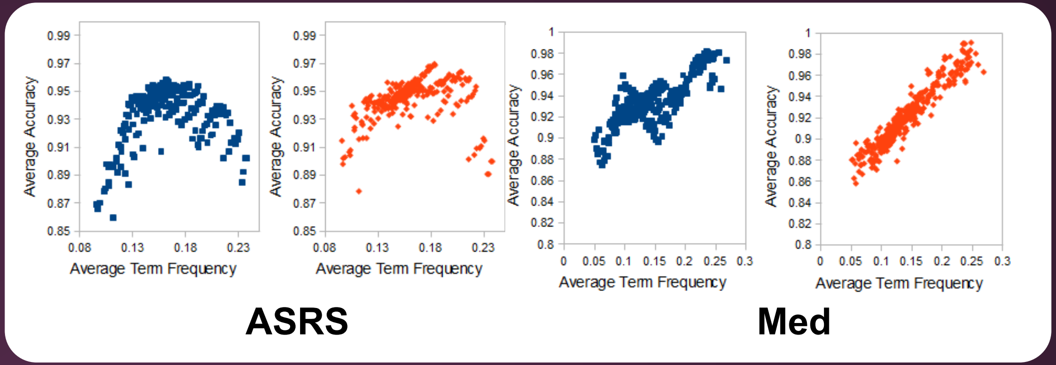
Results : GTM > VSM & VSM > GTM

ASRS

Med



Results : GTM & VSM



➢ Generally one approach yielded significantly better results

ASRS

- At higher term frequencies → worst results -- likely because ASRS contained more common terms than other datasets

Med

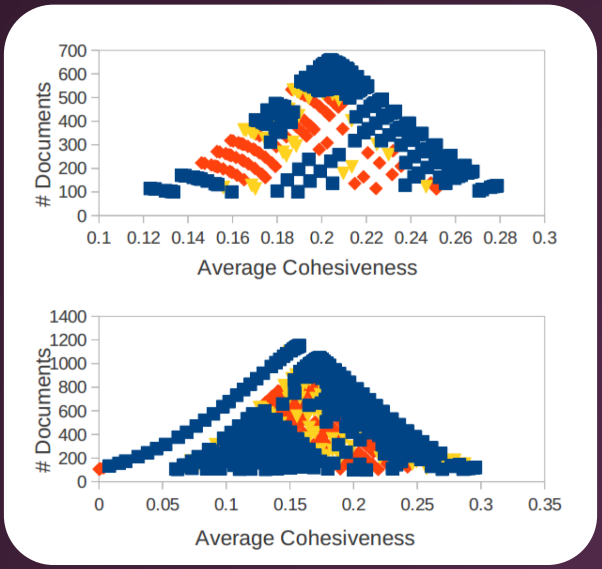
- Higher term frequency → higher accuracies, similar to BHLs

Comparison of Trigram & VSM Approaches: based on term frequency

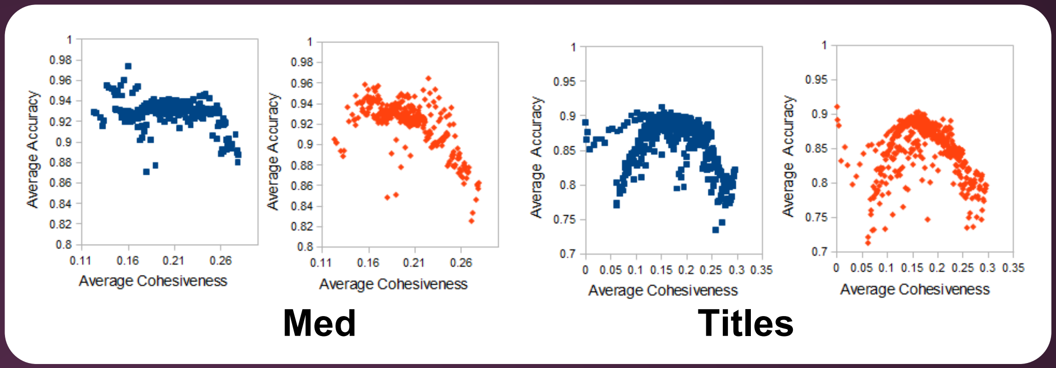
Results : GTM > VSM & VSM > GTM

Med

Titles



Results : GTM & VSM



➢ Generally one approach yielded significantly results, again

Med & Titles

- Cohesiveness played a larger role on these smaller documents

In general...

- Higher cohesion $\simeq$ shorter documents = harder to classify
- More refinement of this measure is needed

Comparison of Trigram & VSM Approaches: based on cohesiveness

Limitations

Data limitations:

- Limited to a 3-5 different dataset (important in result)
- Data length (affect program running speed)
- Data category limitation (2-5), single category assumed
- Data size (300-1000)

Evaluation Limitations:

- Truth based on categories, not similarities
- Observance of correlation with document attributes
- Little regard for actual values of f-measure/accuracy
- Limited attributes, single attribute results

Conclusions

Experimental Results:

- Presented findings of how one approach significantly does better depending on:
  - Genre (dataset source)
  - Document length
  - Term frequency
  - Cohesiveness

Future Work

Overcoming Limitations:

- Investigating more documents
  - More categories, different types, different lengths

Refining observation causation:

- Finding impact of document attributes on results...

References

[1] Islam, Aminul, Evangelos E. Milios, and Keselj Vlado. "Comparing Word Relatedness Measures Based on Google N-grams." *COLING (Posters)* (2012): 495-506. Web. 7 May 2013. <https://web.cs.dal.ca/~eem/cvWeb/pubs/2012-Aminul-Coling.pdf>.

[2] Oza, Nikunj. "SIAM 2007 Text Mining Competition Dataset." *DASHlink*. NASA, 22 Sept. 2010. Web. 31 May 2013.

[3] "Biodiversity Heritage Library." *Biodiversity Heritage Library*. N.p., n.d. Web. 07 Aug. 2013.

[4] Inkpen, Diana. "Solution to the Example." *CSI4107: Information Retrieval and the Internet*. UOttawa, 13 Jan. 2013. Web. 30 June 2013. <http://www.site.uottawa.ca/~diana/csi4107/>.

[5] Soboroff, Ian. "IR Models: The Vector Space Model." *Information Retrieval*. UMBC, 1 Oct. 2002. Web. 7 Aug. 2013. <http://www.csee.umbc.edu/~ian/irF02/lectures/07Models-VSM.pdf>.

[6] Arguello, Jaime. "Vector Space Model." *INLS 509: Information Retrieval*. UNC, 19 Sept. 2011. Web. 7 Aug. 2013. <http://ils.unc.edu/courses/2011_fall/inls509_001/lectures/07-Vector%20Space%20Model.pdf>.

[7] Islam, Aminul, Evangelos Milios, and Vlado Keselj. "Text Similarity Using Google Tri-grams." *Lecture Notes in Computer Science* 7310 (2012): 312-17. Web. 7 May 2013. <http://link.springer.com/chapter/10.1007%2F978-3-642-30353-1_29>.

Thank you for listening!

QUESTIONS?

Slides


1
2
Introduction
3
Motivation
4
Contributions
5
Related Work
6
Methodology
7
Datasets
8
Datasets
9
Datasets: ASRS
10
Datasets: Med
11
Datasets: BHL
12
13
Approaches: VSM
14
Approaches: GTM
15
Approaches: GTM
16
17
Evaluation: Setup
18
kNN-Classification
19
20
21
22
Limitations
23
Conclusions
24
Future Work

Attribute Definitions

- **Document length:**

  # words in the document

- **Term frequency:**

$$\frac{\sum_{\text{each document word}} \text{frequency of that word in dataset}}{\text{\# document words}}$$

- **Cohesiveness**

$$\frac{\sum_{\text{each document word}} \text{word similarity between word and next}}{\text{\# document words - 1}}$$

Results Summarization

Dataset	Limits Min.	Max.	Int.	GTM ? VSM >	<	no diff.	Attr. Correlation GTM		VSM	
Word Count:										
ASRS	6	302	8	36.6	41.7	21.7	Pl	**0.662**	Np	0.366
Med	2	100	2	62.2	26.0	11.8	Pp	0.531	Pp	**0.603**
BHL Titles	0	36	2	67.5	14.2	18.3	Pp	0.004	Pp	**0.031**
BHL Intro	53	539	9	0.0	99.0	0.1	Nl	0.335	Nl	**0.625**
Term Frequency:										
ASRS	0.04	0.36	0.01	17.5	57.3	25.2	Np	**0.713**	Np	0.561
Med	0.01	0.52	0.01	68.0	23.6	8.4	Pl	0.721	Pl	**0.931**
BHL Titles	0.00	1.00	0.05	63.8	30.7	5.5	Np	**0.604**	Np	0.578
BHL Intro	0.03	0.21	0.01	1.0	91.0	8.0	Pp	**0.859**	Pp	0.834
Cohesion:										
ASRS	0.15	0.30	0.01	20.8	65.3	13.9	Np	**0.889**	Np	0.882
Med	0.00	0.37	0.01	74.1	17.3	8.6	Np	0.276	Np	**0.620**
BHL Titles	0.00	0.45	0.01	79.5	9.3	11.2	Np	**0.517**	Np	0.470
BHL Intro	0.05	0.35	0.01	0.0	99.3	0.0	Np	**0.743**	Np	0.719

[7] Islam, A., Milios, E., Vlado K.: Text Similarity using Google Tri-grams. In: Advances in Artificial Intelligence; Lecture Notes in Computer Science. ‘12.

Approaches: Trigram Model

Google Trigram Model - Word Similarity:

$\Rightarrow$ word_sim(,)

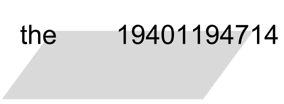
Consider trigram frequencies w.r.t. all pair’s unigram frequencies

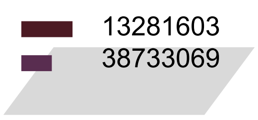
$$
= \begin{cases}
\dfrac{\log \dfrac{\mu(w_a,n_1,w_b,n_2)C^2}{c(w_a)c(w_b)\min(c(w_a),c(w_b))}}{-2\times\log\dfrac{\min(c(w_a),c(w_b))}{C}} & \text{if } \dfrac{\mu(w_a,n_1,w_b,n_2)C^2}{c(w_a)c(w_b)\min(c(w_a),c(w_b))} > 1 \\[2ex]
\dfrac{\log 1.01}{-2\times\log\dfrac{\min(c(w_a),c(w_b))}{C}} & \text{if } \dfrac{\mu(w_a,n_1,w_b,n_2)C^2}{c(w_a)c(w_b)\min(c(w_a),c(w_b))} <= 1 \\[2ex]
0 & \text{if } \mu(w_a,n_1,w_b,n_2) = 0
\end{cases}
$$

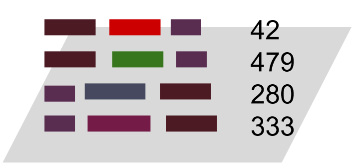
Maximum frequency of unigrams

Frequency of word (w_a or w_b) in unigrams

Mean frequency of n_1 trigrams that start with w_a & end with w_b and n_2 trigrams that start with w_b & end with w_a







[5] http://www.csee.umbc.edu/~ian/irF02/lectures/07Models-VSM.pdf
[6] http://ils.unc.edu/courses/2011_fall/inls509_001/lectures/07-Vector%20Space%20Model.pdf

Approaches: VSM

Advantages:

- Very commonly used, works well, simple
- Weighting is based off importance in dataset
- Counts for partial matches

Disadvantages:

- Representation suffers when #words = too long / short
- Requires a "large" dataset to calculate meaningful IDF
- Dependent on common words being present within same category

[7] Islam, A., Milios, E., Vlado K.: Text Similarity using Google Tri-grams. In: Advances in Artificial Intelligence; Lecture Notes in Computer Science. '12.

Approaches: Trigram Model

Advantages:

- Partial matching via Google n-gram word similarity
- Can simply calculate the relatedness between two documents

Disadvantages:

- Dependent on Google n-gram coverage (relative to testing dataset) → Special words problem
- Requires large corpus (Google n-grams) to calculate relatedness

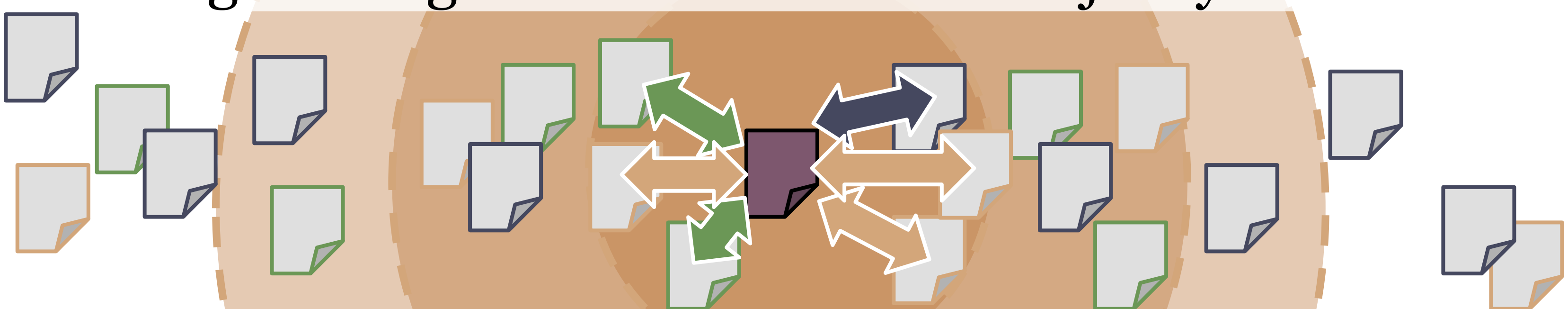
http://www.clker.com/clipart-176644.html

Evaluations: kNN-Classification

How to Calculate Accuracy:

(Executed using each partition = testing set; then average **10** resultant accuracies → **30**

1. Consider different k from [1,$\sqrt{\text{\# testing set}}$]
2. For each document in testing set, assign class based on training set neighbours' *normalized* majority class



3. Calculate accuracy:

   ⇒ # correctly assigned docs / # testing docs

4. Select k where mean accuracy is highest → accuracy