

Geographical Classification of Documents Using Evidence from Wikipedia

Rafael Odon de Alencar (odon.rafael@gmail.com)

Clodoveu Augusto Davis Jr. (clodoveu@dcc.ufmg.br)

Marcos André Gonçalves (mgoncalv@dcc.ufmg.br)

Universidade Federal de Minas Gerais, Brazil

GIR'10, 18-19th Feb. 2010, Zurich, Switzerland



UFMG - ICEx
DEPARTAMENTO DE CIÊNCIA DA
COMPUTAÇÃO



Introduction / Motivation

(common to most of GIR 2010)

- Geography-related terms are often used in Web search queries
- Many user activities on the Web are directly related to the user's location
- It is important to conceive applications that take into consideration this intention

Introduction

- Recent work has suggested identifying the geographic context of documents
 - Association of Web pages to places
- Advances can enhance current information retrieval mechanisms
 - Allow people to perform *local search*
 - Enable geographically-focused advertising
 - Develop novel ranking strategies

Introduction

- Identification of the geographic context of a Web document:
 - Inferred by the location of its Web server (GeoIP)
 - Inferred by the location of its visitors and of adjacent pages in the Web graph
 - Determined by analyzing the document's textual content

Introduction

- Identification of the geographic context of a document:
 - Inferred by the location of its Web server (GeoIP)
 - Inferred by the location of its visitors and of adjacent pages in the Web graph.
 - **Determined by analyzing the document's textual content**

Introduction

- In previous work, our group has developed means to recognize direct and indirect evidence of location, using an extraction ontology
 - Addresses
 - Postal codes
 - Telephone numbers and area codes
 - Positioning expressions:

<place of interest> <location expression> <landmark>

Hotel CLOSE TO Convention Center

Introduction

- However, not all pages include unambiguous and easily recognizable evidence
- This work looks at other types of textual evidence
 - Terms and expressions semantically related to a location
 - Not necessarily other place names

— **Minas Gerais** é uma das 27 [unidades federativas](#) do [Brasil](#), sendo a quarta maior em extensão territorial, que é de 586.528 km², equivalente à da [França](#). Localiza-se no [Sudeste](#) e limita-se a sul e sudoeste com [São Paulo](#), a oeste com o [Mato Grosso do Sul](#) e a noroeste com [Goiás](#), incluindo uma pequena divisa com o [Distrito Federal](#), a leste com o [Espírito Santo](#), a sudeste com o [Rio de Janeiro](#) e a norte e nordeste com a [Bahia](#). O atual governador do estado é [Aécio Neves](#). Linguisticamente, o nome Minas Gerais dentro de frases não é acompanhado de artigo definido, como acontece com os estados de Mato Grosso, de Goiás e de Mato Grosso do Sul.

O estado é o segundo mais populoso do [Brasil](#), com pouco mais de 20 milhões de habitantes^[3]. Sua capital é a cidade de [Belo Horizonte](#), que reúne em sua [região metropolitana](#) cerca de cinco milhões de habitantes.

Minas Gerais possui o terceiro maior [Produto Interno Bruto](#) do [Brasil](#), superado apenas pelos estados de [São Paulo](#) e [Rio de Janeiro](#), embora em um importante indicador de capacidade econômica, a arrecadação de [ICMS](#), Minas tenha superado o [Rio de Janeiro](#) na classificação nacional.

Minas Gerais é muito importante também sob o aspecto [histórico](#): cidades erguidas durante o ciclo do [ouro](#) no [século XVIII](#) consolidaram a colonização do interior do país e estão espalhadas por todo o estado. Alguns eventos marcantes da história brasileira, como a [Inconfidência Mineira](#), a [Revolução de 1930](#) e o [Golpe Militar de 1964](#) foram arquitetados em Minas Gerais.



— **Minas Gerais** é uma das 27 [unidades federativas](#) do [Brasil](#), sendo a quarta maior em extensão territorial, que é de 586.528 km², equivalente à da [França](#). Localiza-se no [Sudeste](#) e limita-se a sul e sudoeste com [São Paulo](#), a oeste com o [Mato Grosso do Sul](#) e a noroeste com [Goiás](#), incluindo uma pequena divisa com o [Distrito Federal](#), a leste com o [Espírito Santo](#) e a sudeste com o [Rio de Janeiro](#) e a norte e nordeste com a [Bahia](#). O atual governador do estado é [Aécio Neves](#). Linguisticamente, o nome Minas Gerais dentro de frases não é acompanhado de artigo definido, como acontece com os estados de Mato Grosso, de Goiás e de Mato Grosso do Sul.

O estado é o segundo mais populoso do [Brasil](#), com pouco mais de 20 milhões de habitantes^[3]. Sua capital é a cidade de [Belo Horizonte](#) que reúne em sua [região metropolitana](#) cerca de cinco milhões de habitantes.

Minas Gerais possui o terceiro maior [Produto Interno Bruto](#) do [Brasil](#), superado apenas pelos estados de [São Paulo](#) e [Rio de Janeiro](#), embora em um importante indicador de capacidade econômica, a arrecadação de [ICMS](#), Minas tenha superado o [Rio de Janeiro](#) na classificação nacional.

Minas Gerais é muito importante também sob o aspecto [histórico](#): cidades erguidas durante o ciclo do [ouro](#) no [século XVIII](#) consolidaram a colonização do interior do país e estão espalhadas por todo o estado. Alguns eventos marcantes da história brasileira, como a [Inconfidência Mineira](#), a [Revolução de 1930](#) e o [Golpe Militar de 1964](#) foram arquitetados em Minas Gerais.



— **Minas Gerais** é uma das 27 [unidades federativas](#) do [Brasil](#), sendo a quarta maior em extensão territorial, que é de 586.528 km², equivalente à da [França](#). Localiza-se no [Sudeste](#) e limita-se a sul e sudoeste com [São Paulo](#), a oeste com o [Mato Grosso do Sul](#) e a noroeste com [Goiás](#), incluindo uma pequena divisa com o [Distrito Federal](#), a leste com o [Espírito Santo](#) e a sudeste com o [Rio de Janeiro](#) e a norte e nordeste com a [Bahia](#). O atual governador do estado é [Aécio Neves](#). Linguisticamente, o nome Minas Gerais dentro de frases não é acompanhado de artigo definido, como acontece com os estados de Mato Grosso, de Goiás e de Mato Grosso do Sul.

O estado é o segundo mais populoso do [Brasil](#), com pouco mais de 20 milhões de habitantes^[3]. Sua capital é a cidade de [Belo Horizonte](#) que reúne em sua [região metropolitana](#) cerca de cinco milhões de habitantes.

Minas Gerais possui o terceiro maior [Produto Interno Bruto](#) do [Brasil](#), superado apenas pelos estados de [São Paulo](#) e [Rio de Janeiro](#), embora em um importante indicador de capacidade econômica, a arrecadação de [ICMS](#), Minas tenha superado o [Rio de Janeiro](#) na classificação nacional.

Minas Gerais é muito importante também sob o aspecto [histórico](#): cidades erguidas durante o ciclo do [ouro](#) no [século XVIII](#) consolidaram a colonização do interior do país e estão espalhadas por todo o estado. Alguns eventos marcantes da história brasileira, como a [Inconfidência Mineira](#), a [Revolução de 1930](#) e o [Golpe Militar de 1964](#) foram arquitetados em Minas Gerais.



— **Minas Gerais** é uma das 27 [unidades federativas](#) do [Brasil](#), sendo a quarta maior em extensão territorial, que é de 586.528 km², equivalente à da [França](#). Localiza-se no [Sudeste](#) e limita-se a sul e sudoeste com [São Paulo](#), a oeste com o [Mato Grosso do Sul](#) e a noroeste com [Goiás](#), incluindo uma pequena divisa com o [Distrito Federal](#), a leste com o [Espírito Santo](#) e a sudeste com o [Rio de Janeiro](#) e a norte e nordeste com a [Bahia](#). O atual governador do estado é [Aécio Neves](#). Linguisticamente, o nome Minas Gerais dentro de frases não é acompanhado de artigo definido, como acontece com os estados de Mato Grosso, de Goiás e de Mato Grosso do Sul.

O estado é o segundo mais populoso do [Brasil](#), com pouco mais de 20 milhões de habitantes^[3]. Sua capital é a cidade de [Belo Horizonte](#) que reúne em sua [região metropolitana](#) cerca de cinco milhões de habitantes.

Minas Gerais possui o terceiro maior [Produto Interno Bruto](#) do [Brasil](#), superado apenas pelos estados de [São Paulo](#) e [Rio de Janeiro](#), embora em um importante indicador de capacidade econômica, a arrecadação de [ICMS](#), Minas tenha superado o [Rio de Janeiro](#) na classificação nacional.

Minas Gerais é muito importante também sob o aspecto [histórico](#): cidades erguidas durante o ciclo do [ouro](#) no [século XVIII](#) consolidaram a colonização do interior do país e estão espalhadas por todo o estado. Alguns eventos marcantes da história brasileira, como a [Inconfidência Mineira](#), a [Revolução de 1930](#) e o [Golpe Militar de 1964](#) foram arquitetados em Minas Gerais.





Museu da Inconfidência em Ouro Preto.

Literatura

A [literatura](#) floresce em Minas Gerais já na época do ouro, através principalmente de [Alvarenga Peixoto](#) e [Cláudio Manoel da Costa](#).

Modernamente, Minas Gerais deu sua contribuição à literatura brasileira através de es

- [Autran Dourado](#) - natural de Patos de Minas, no Alto Paranaíba, autor de *Uma Vida er*
- [João Guimarães Rosa](#) - autor de *Grande Sertão: Veredas* e *Sagarana* entre outros.
- [Fernando Sabino](#) - autor de *O Encontro Marcado*, *O Grande Mentecapto* e diversos livros de contos e crônicas.
- [Carlos Drummond de Andrade](#) - um de seus poemas mais conhecidos é "*Itabira é só um retrato na parede, mas como dói.*", a ele estar morando há muito no [Rio](#).
- [Adélia Prado](#) - autora de *O coração disparado*, *O homem da mão seca* e outras obras na poesia e em prosa.
- [Murilo Rubião](#) - autor de contos fantásticos, como *O Pirotécnico Zacarias* e *O Ex-Mágico*.
- [Rubem Fonseca](#) - autor de *Agosto*, *Bufo & Spalanzani*, entre outros.
- [Ziraldo](#) - humorista e criador do [Menino Maluquinho](#).
- [Pedro Nava](#) - autor de *Baú de Ossos*, *Chão de Ferro*, *Beira-mar*, entre outros.
- [Murilo Mendes](#) - poeta, autor de importantes obras do [Modernismo brasileiro](#).
- [Darcy Ribeiro](#) - autor de romances como *Maíra*, *Migo* e de diversas obras voltadas à Educação, Antropologia e Etnologia.
- [Roberto Drummond](#) - autor de *Hilda Furacão*, *O Cheiro de Deus*, *Hitler manda lembranças*, dentre outros.
- [Mário Palmério](#) - autor de *Vila dos Confins* e *Chapadão do Bugre*.



Our Proposal

- Use the Wikipedia as a semantic network, composed by its entries (nodes) and links (arcs), to gather textual geographic evidence for places

Our Proposal

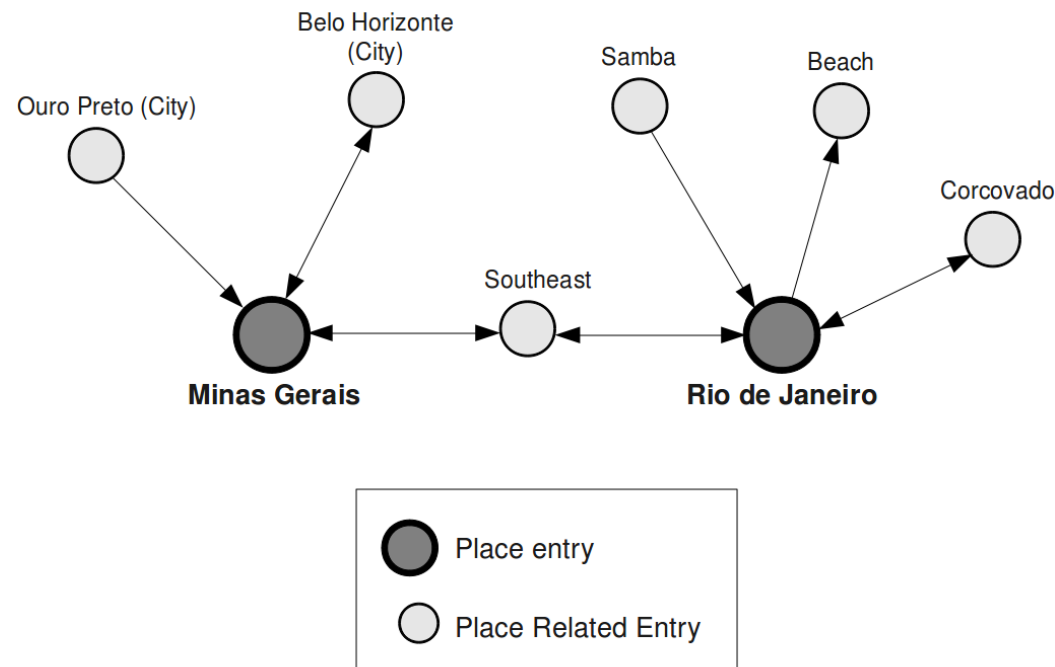
- This work intends to demonstrate that such evidence is valid using classification experiments
 - Classes: a subset of Brazilian states, considered as single labels
 - Database: a set of articles from the “local news” section of newspapers
- We don't intend to propose a definitive geographic classification model

Geographic Evidence from Wikipedia

- Start with a set of places
- Find the Wikipedia entry for each place
- Collect the titles of *inlinks* and of *outlinks*
 - Titles of entries are used as terms for IR
- Use weights to indicate how frequent (how important) a term is
- Organize such information as evidence for a classifier

Geographic Evidence from Wikipedia

- Consider a set of places and its adjacent entries (links) in Wikipedia



Geographic Evidence from Wikipedia

- Each place has a list of inlinks and outlinks
- Weights are used to inform the discriminative value of each term

Minas Gerais inlinks	
Article	w(t)
Ouro Preto (city)	1
Belo Horizonte (city)	1
Southeast	0.25

Rio de Janeiro inlinks	
Article	w(t)
Samba	1
Corcovado	1
Southeast	0.25

Minas Gerais outlinks	
Article	w(t)
Belo Horizonte (city)	1
Southeast	0.25

Rio de Janeiro outlinks	
Article	w(t)
Beach	1
Corcovado	1
Southeast	0.25

Geographic Evidence from Wikipedia

- The weight of a term t is based on its adjacency to the considered set of m places
 - More “exclusive” terms have a weight close to 1.0
 - More “popular” terms have a weight close to 0.0

$$w(t) = \left(1 - \frac{adj(t) - 1}{m} \right)^2$$

Geographic Evidence from Wikipedia

- Classification: we find occurrences of entry titles in documents

Document 1

Our company
has offices in
Belo Horizonte
and **Ouro**
Preto...

Document 2

This year's **samba**
festival will also
occur in other
Brazilian **southeast**
state capitals...

Geographic Evidence from Wikipedia

- Considering the occurrences found, we use weighted sums to describe the relationship of a document to places from the set

$$S^{in}(p_i, d_j) = \sum_{l=1}^{|in|} w(t_l) \times \text{Frequency}(t_l, d_j) \quad S^{out}(p_i, d_j) = \sum_{l=1}^{|out|} w(t_l) \times \text{Frequency}(t_l, d_j)$$

Document 1	Minas Gerais		Rio de Janeiro		Class
	in	out	in	out	
Document 1	2	1	0	0	Minas Gerais
Document 2	0.25	0.25	1.25	0.25	Rio de Janeiro

Geographic Evidence from Wikipedia

- Some improvements were done, in order to get a richer description
 - Use separate sums for every level of importance
 - This can lead a classifier to better “understand” the relationship between the text and the places

Document 1	Minas Gerais				Rio de Janeiro				Class
	in ₁	in ₂	out ₁	out ₂	in ₁	in ₂	out ₁	out ₂	
Document 1	2	0	1	0	0	0	0	0	Minas Gerais
Document 2	0	0.25	0	0.25	1	0.25	0	0.25	Rio de Janeiro

Document collection

- We classified documents associated to Brazilian states
 - We considered a subset of 8 from the 27 Brazilian states
 - We extracted 831 articles from 8 different local news sections
 - We read the titles of each article to be sure they were indeed related to the respective state

Document collection

- Only the article title and its body were extracted (no structure was preserved)
- The text was pre-processed:
 - Stemming: words reduced to radical form
 - Stopwords removal: ignore conjunctions, prepositions, punctuations and other inexpressive words.

Document collection

Brazilian State	News Web Site	Local News Section Name	Access Date	# Docs
Minas Gerais (MG)	www.uai.com.br	Minas	09/11/09	104
São Paulo (SP)	g1.globo.com	São Paulo	09/11/09	101
Rio de Janeiro (RJ)	g1.globo.com	Rio	09/11/09	102
Acre (AC)	www.ac24horas.com.br	Acre	10/11/09	103
Amazonas (AM)	www.amazonasagora.com.br	Municípios	25/11/09	103
Bahia (BA)	ibahia.globo.com	Bahia	30/11/09	105
Santa Catarina (SC)	www.folhanorte.com.br	All	01/12/09	109
Pernambuco (PE)	www.diariodepernambuco.com.br	Vida Urbana	01/12/09	104
Total				831

Evaluation

- We chose the Multinomial Naïve Bayes Classifier to perform our tests
 - Features represent the frequency of terms
 - Ignores the position of the terms in texts
 - Considers features to be independent (*naïve assumption*)
 - In practice this simplifies the learning process
 - Adjusts a model based on the probability of a class to generate an instance considering the given examples

Evaluation

- N-fold cross validation was used for tests
 - The dataset is divided in N parts
 - Every part is used as a test set for training with the other N-1 parts
 - Every instance is guaranteed to be used both for test and training
 - Success rate is obtained from the whole dataset
- All tests performed using Weka 3.6.1

Evaluation

- TF-IDF measurements of a bag-of-words representation of documents was used as a baseline for our evaluation

- *Bag-of-words*: reduces documents to lists of terms
- TF-IDF: gives us terms frequencies, normalized by the document length and term popularity in the collection.

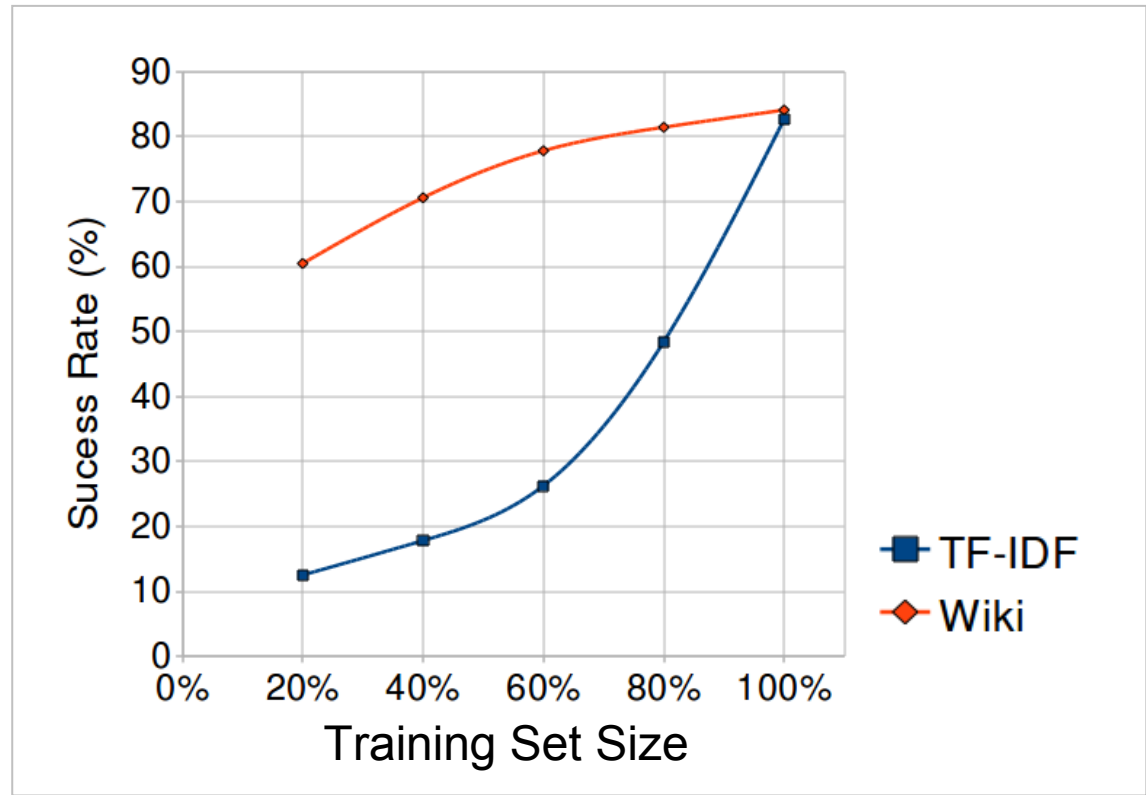
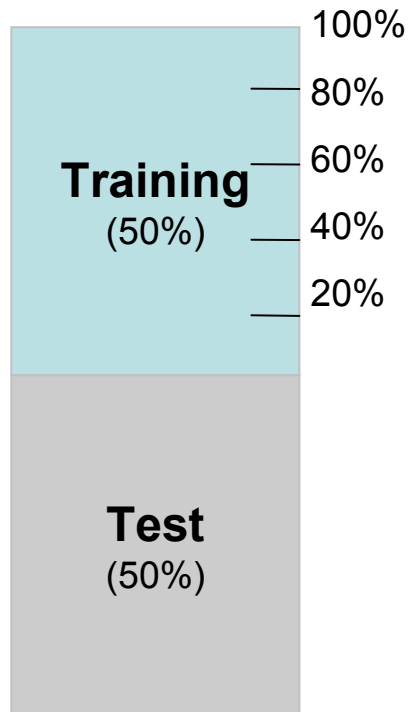
$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

$$(tfidf)_{i,j} = tf_{i,j} \times idf_i$$

Evaluation

Results – Success rate for different training set sizes



Technique	Different training set size(% of a fixed subset)				
	100%	80%	60%	40%	20%
TF-IDF	82.65	48.43	26.27	17.83	12.53
Wiki	84.1	81.45	77.83	70.6	60.48

Evaluation

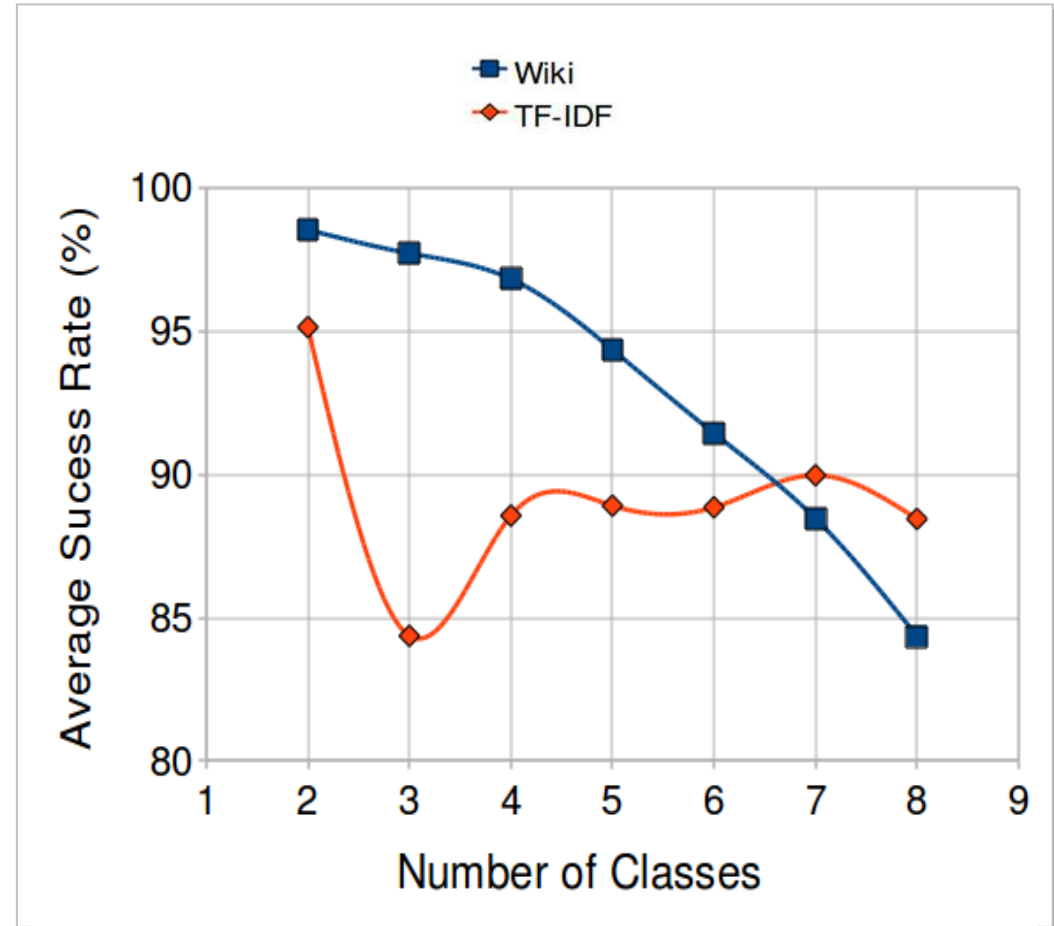
Results – Success rate for different number of classes

TF-IDF

Random Seed Value	Number of Classes						
	2	3	4	5	6	7	8
1	97.09	84.69	88.56	89.69	88.69	89.29	88.81
2	96.12	85.34	88.56	88.52	89.5	89.97	88.69
3	94.66	84.36	88.32	88.52	88.85	89.7	87.73
4	95.15	83.06	87.83	89.3	90.47	90.11	88.45
5	94.66	83.71	88.81	88.91	88.69	90.11	88.33
Std. Dev.	1.05	0.88	0.37	0.51	0.76	0.35	0.42
Average	95.15	84.36	88.56	88.91	88.85	89.97	88.45

Wiki

Random Seed Value	Number of Classes						
	2	3	4	5	6	7	8
1	98.06	97.72	96.83	94.16	91.76	88.74	84.48
2	98.54	97.72	96.83	94.55	91.11	88.6	84.36
3	98.54	97.72	96.83	94.75	91.44	88.46	84
4	98.54	97.72	96.59	94.36	91.6	87.91	83.87
5	98.06	97.72	96.59	94.36	90.79	88.32	84.48
Std. Dev.	0.27	0	0.13	0.22	0.39	0.32	0.28
Average	98.54	97.72	96.83	94.36	91.44	88.46	84.36



Evaluation

Results – Effects of removing place names

- Our hypothesis: TF-IDF of bag-of-words classification has a non-geographic bias
 - Other irrelevant terms are represented by the features
- We defined 100 place names to be removed from the documents in order to check the impact on precision
 - State names, abbreviations, important city names, and others
 - More than 35,000 removals

Evaluation

Results – Effects of removing place names

- We classified the dataset before and after the place names removal for comparison
- 10-fold cross validation was used
- Impact in precision:
 - Wikipedia Model: more than **30% of loss**
 - TF-IDF Bag-of-words model: about **6% of loss**

Conclusion

- Wikipedia model
 - Operates very well with less training
 - Adding more classes decreases its precision
 - Is sensible to geographic evidence in the text
- TF-IDF bag-of-words model
 - Captures non-geographic detail from the training documents

Conclusion

- Future work includes:
 - Generate a much larger collection, from more sources, and considering the 27 states
 - Mix types of places: states, cities, countries, etc.
 - Improve the matching of entry titles in documents by considering alternatives or synonyms for them
 - Consider multi-label classification

Direções atuais:

- Classificação não é o caminho ideal, apenas um experimento interessante.
- Grande parte do sucesso está na identificação correta do “assunto” do texto:
 - Keywords extraction
 - Topic Indexing
 - Reproduzir o mecanismo de Wikificação do Milne & Witten 2008
- Construir coleção de documentos com aspectos hierárquicos, multi-classe, com ou sem contexto definido.

Geographical Classification of Documents Using Evidence from Wikipedia

Rafael Odon de Alencar (odon.rafael@gmail.com)

Clodoveu Augusto Davis Jr. (clodoveu@dcc.ufmg.br)

Marcos André Gonçalves (mgoncalv@dcc.ufmg.br)

Universidade Federal de Minas Gerais, Brazil

GIR'10, 18-19th Feb. 2010, Zurich, Switzerland



UFMG - ICEx
DEPARTAMENTO DE CIÊNCIA DA
COMPUTAÇÃO

