EXTRAINDO TEMA DE DOCUMENTOS DIPLOMÁTICOS

Alice Duarte Scarpa

29 de Outubro de 2015

INTRODUÇÃO

Os Dados

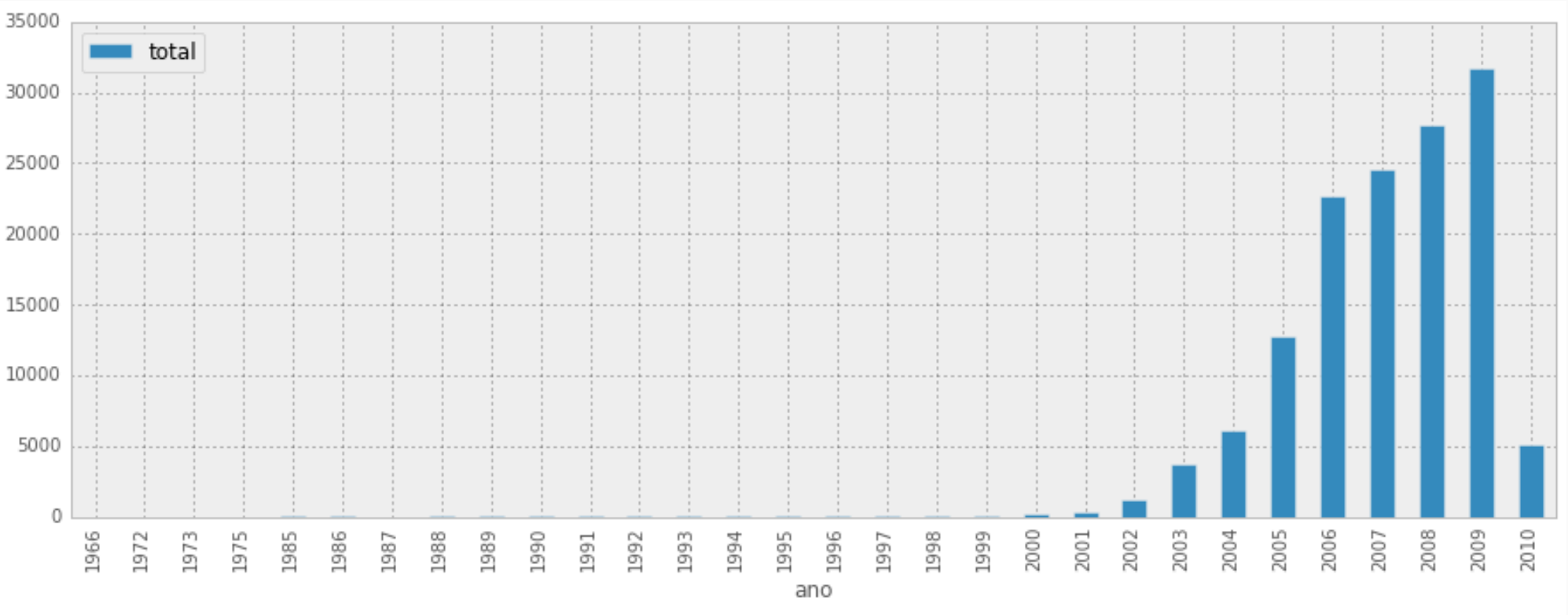
Neste trabalho vamos explorar dados do cablegate, que são cabos diplomáticos entre embaixadas dos Estados Unidos que foram divulgados como parte do WikiLeaks.

Os Dados

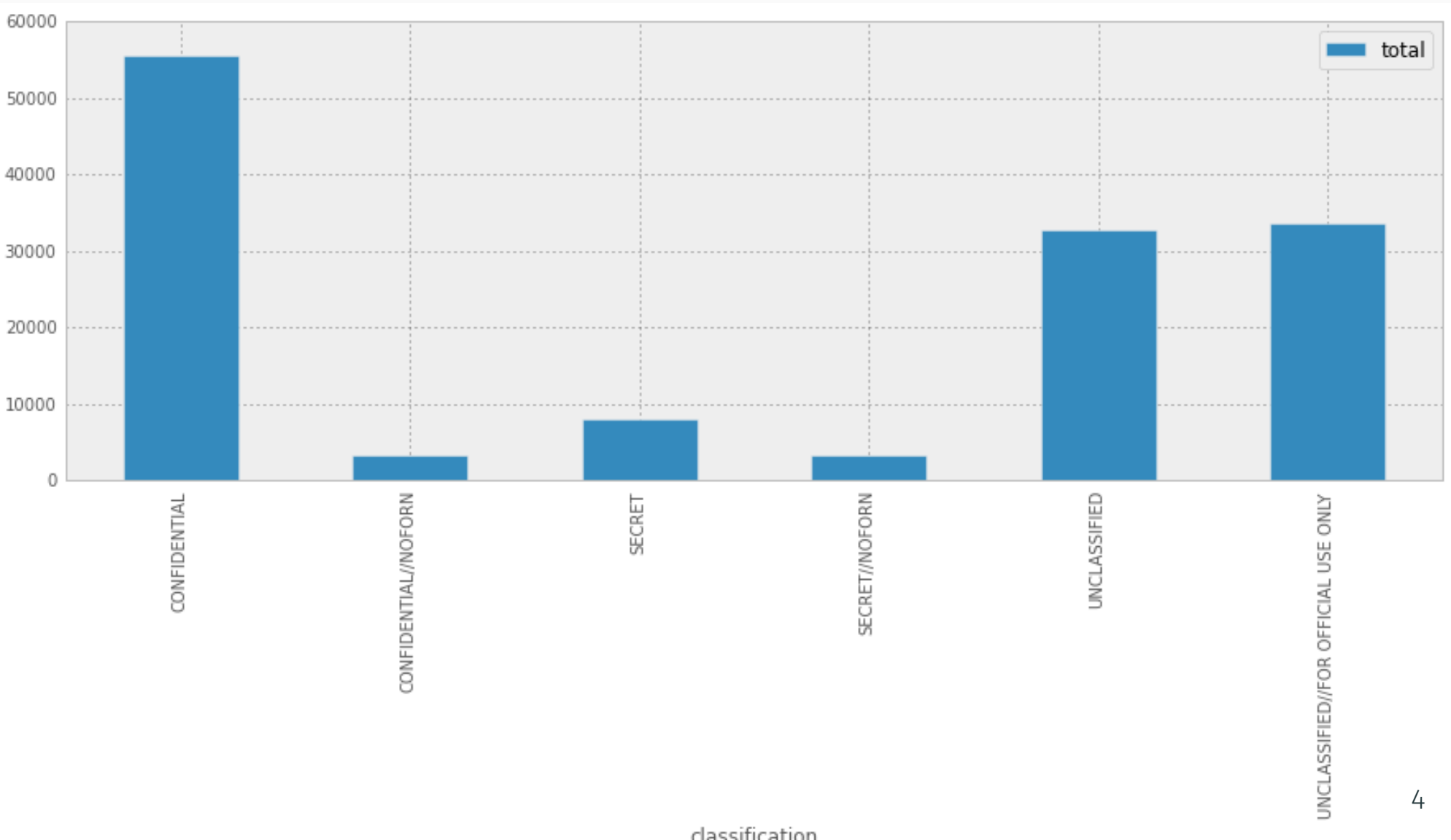
Neste trabalho vamos explorar dados do cablegate, que são cabos diplomáticos entre embaixadas dos Estados Unidos que foram divulgados como parte do WikiLeaks.

São 250 mil documentos, de 1966 a 2010 e diferentes níveis de confidencialidade. A grande maioria dos documentos é de 2003 a 2010.

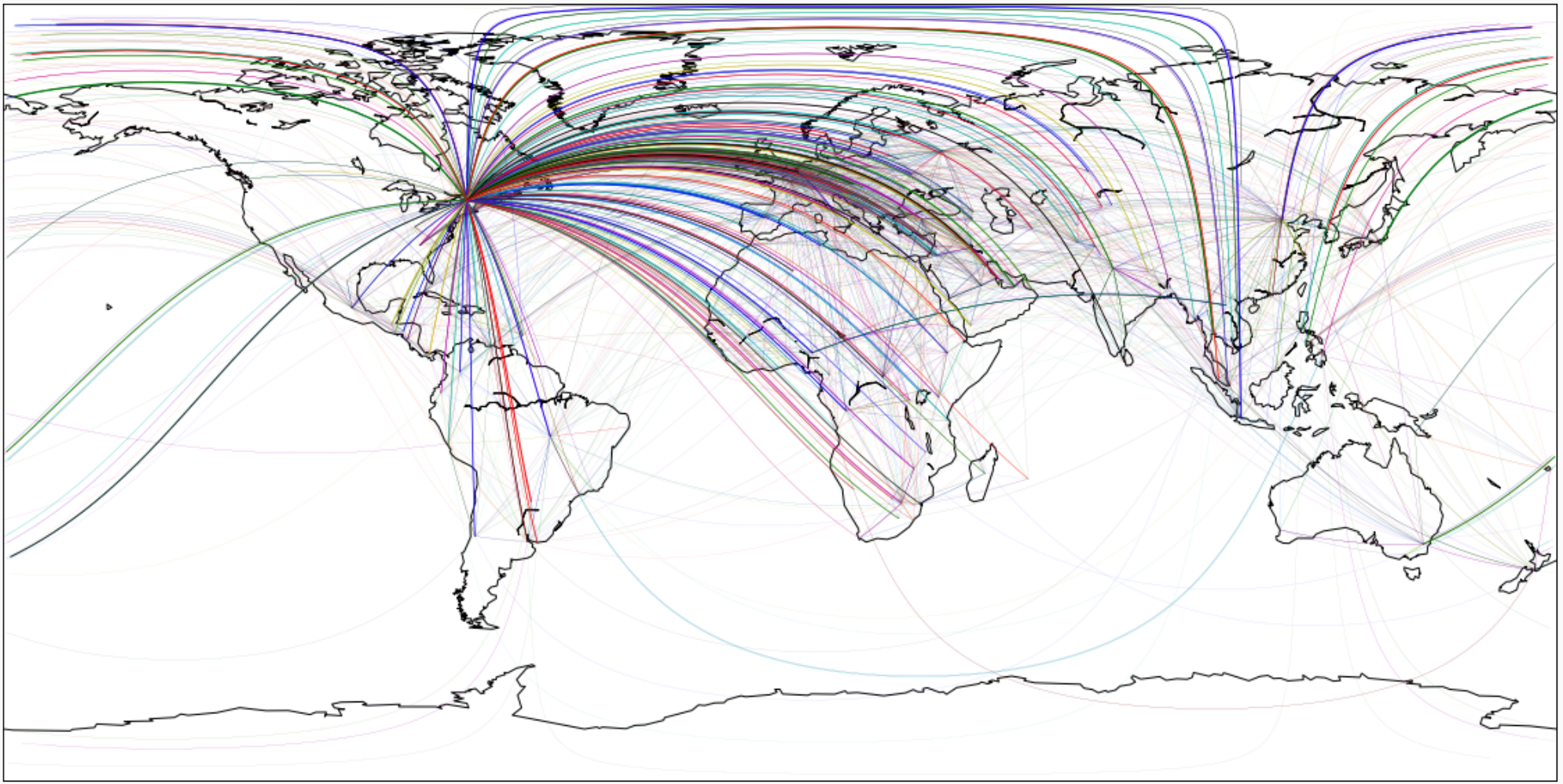
Os Dados


total
35000
30000
25000
20000
15000
10000
5000
0
1966
1972
1973
1975
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
ano

Os Dados


60000
50000
40000
30000
20000
10000
0
total
CONFIDENTIAL
CONFIDENTIAL//NOFORN
SECRET
SECRET//NOFORN
UNCLASSIFIED
UNCLASSIFIED//FOR OFFICIAL USE ONLY
classification

Os Dados



OBJETIVOS

Objetivos

- Criar uma base de dados explorável

Objetivos

- Criar uma base de dados explorável
- Comparar diferentes técnicas existentes de extração de tema

Objetivos

- Criar uma base de dados explorável
- Comparar diferentes técnicas existentes de extração de tema
- Explorar características dos documentos para adaptar alguma técnica existente

TÉCNICAS

Mecânicas para pré-processamento

Bag of Words Considera o texto apenas como o conjunto de palavras que nele aparece e a quantidade de vezes que cada palavra aparece.

Mecânicas para pré-processamento

Bag of Words Considera o texto apenas como o conjunto de palavras que nele aparece e a quantidade de vezes que cada palavra aparece.

N-gramas Como o Bag of Words, mas olha para expressões de n palavras.

Mecânicas para pré-processamento

Bag of Words Considera o texto apenas como o conjunto de palavras que nele aparece e a quantidade de vezes que cada palavra aparece.

N-gramas Como o Bag of Words, mas olha para expressões de n palavras.

Noun Phrases Procura o sujeito e o objeto de cada oração.

Técnicas de Extração de Tema

N.E.E. (Named Entities Extraction) Tenta responder: Quem? Como? Onde?

Técnicas de Extração de Tema

N.E.E. (Named Entities Extraction) Tenta responder: Quem? Como? Onde?

Lexical Chaining Procura cadeias semânticas no texto usando sinônimos, antônimos e repetições

Técnicas de Extração de Tema

N.E.E. (Named Entities Extraction) Tenta responder: Quem? Como? Onde?

Lexical Chaining Procura cadeias semânticas no texto usando sinônimos, antônimos e repetições

Candidatos a tema A partir das Part-Of-Speech Tags, determina alguns candidatos a tema e dá uma nota para cada um deles

APLICAÇÕES

Aplicações

- Extrair conhecimento a partir de tendências nos temas dos textos

Aplicações

- Extrair conhecimento a partir de tendências nos temas dos textos
- Facilitar o estudo dessa base de dados por pesquisadores de Ciências Sociais

Bibliografia

Experiments with Theme Extraction in Explanatory Texts (1996),
Francois Paradis , Catherine Berrut

Extracting Key Terms From Noisy and Multi-theme Documents (2009),
Maria Grineva, Maxim Grinev, Dmitry Lizorkin

Perguntas?