Text Mining
Aplicación Práctica

Laura Alonso i Alemany

Facultad de Matemática, Astronomía y Física
UNC, Córdoba (Argentina)
http://www.cs.famaf.unc.edu.ar/~laura

SADIO
14 de Marzo de 2008

qué vamos a hacer hoy

1. tarea individual (1.5 h):
   - 1.1 contestar brevemente (aprox 5 líneas) a algunas preguntas cortas sobre conceptos y técnicas de text mining
   - 1.2 elegir 1 de 4 posibles proyectos, y plantear por escrito cómo abordar el problema, argumentando las decisiones
2. *pausa (0.5 h)*
3. tarea grupal (grupos de 2-3 personas) (1 h):
   - 3.1 poner en común las diferentes aproximaciones al problema
   - 3.2 elaborar una argumentación sobre la viabilidad de las diferentes aproximaciones, problemas de implementación, posible evaluación, pros y contras
4. *pausa (0.25 h)*
5. presentación pública de cada grupo (1 h)
6. breve discusión del proyecto individual que desarrollará cada uno en su casa (50% de la evaluación), posiblemente basado en el trabajo desarrollado en clase (0.5 h)
7. valoración del curso (0.25 h)

preguntas cortas

1. mencione las ventajas de bootstrapping en comparación con métodos totalmente supervisados y con métodos levemente supervisados.
2. de qué forma representaría a las palabras como vectores para calcular la semejanza entre ellas? qué características formarían las dimensiones, qué valores se encontrarían en cada dimensión para cada vector?
3. mencione dos formas de alinear textos paralelos, y dos usos posibles de la alineación de textos paralelos.
4. mencione algunas de las barreras que supone la ambigüedad de las palabras para el uso de analizadores automáticos de lenguaje natural.
5. explique la relación entre la adquisición de paráfrasis y la traducción automática estadística.

proyectos

- bootstrapping y enriquecimiento de anotación en texto de biomedicina a partir de un corpus anotado con entidades,
  - aprender un clasificador para identificar entidades en corpus no anotado, teniendo en cuenta que habrá un importante porcentaje de entidades no vistas
  - aprender relaciones entre entidades
- identificar la estructura típica en diálogos telefónicos, para luego identificar posibles diálogos no típicos
- link analysis, social network en un corpus de e-mail
- sentiment analysis en un corpus de e-mail
- identificar entidades y/o identificar fragmentos de texto recurrente, que pueden llegar a formar plantillas de documentos, en un corpus de documentos semi-estructurados (avisos clasificados, artículos de enciclopedia, sesiones parlamentarias)

corpus

- Genia Corpus

  2000 annotated abstracts from MEDLINE, annotated with a subset of the substances and the biological locations involved in reactions of proteins, based on a data model (GENIA ontology) of the biological domain, in XML.

- Corpus OHSUMED
- Corpus EuroParl
- Switchboard Corpus (Penn Treebank Transcriptions)
- Enron email dataset
- wikipedia!

corpus

- Genia Corpus
- Corpus OHSUMED

  348,566 references from MEDLINE, consisting of title, abstract, MeSH indexing terms, author, source, and publication type, incomplete and out-of-date.

- Corpus EuroParl
- Switchboard Corpus (Penn Treebank Transcriptions)
- Enron email dataset
- wikipedia!

corpus

- Genia Corpus
- Corpus OHSUMED
- Corpus EuroParl

  44 million words per language, 11 languages, extracted from the proceedings of the European Parliament, automatically aligned at sentence level.

- Switchboard Corpus (Penn Treebank Transcriptions)
- Enron email dataset
- wikipedia!

corpus

- Genia Corpus
- Corpus OHSUMED
- Corpus EuroParl
- Switchboard Corpus (Penn Treebank Transcriptions)

  2430 spontaneous conversations averaging 6 minutes in length, about 3 million words of text

- Enron email dataset
- wikipedia!

corpus

- Genia Corpus
- Corpus OHSUMED
- Corpus EuroParl
- Switchboard Corpus (Penn Treebank Transcriptions)
- Enron email dataset

  0.5M messages from 50 users, mostly senior management of Enron

- wikipedia!