Sesión 7: Minería de texto

ML-IA: Módulo de Machine learning

Jacob Muñoz-Castro

Universidad de los Andes

24 de noviembre de 2020

Agenda

1. Recap

2. Análisis cuantitativo de textos (QTA)

Clustering

¿Por qué el análisis cuantitativo de texto?

- Metáfora de Justin Grimmer - QTA mejora la lectura
  - Analizar una paja de heno = entender el significado de una oración. *Los humanos somos buenos para eso! Las computadoras...no tanto.*
  - Organizar una pila de heno = describir, clasificar, escalar textos. *Tarea difícil pero no para las computadoras.*

Principios del análisis cuantitativo (Grimmer & Stewart, 2013)

1. Todos los modelos cualitativos están mal, pero algunos son útiles
2. Los modelos cuantitativos para texto amplifican los recursos y son aumentativos para los humanos (capacidad)
3. No hay un método globalmente mejor para el análisis de texto
4. Validar, validar y seguir validando...

Conceptos básicos

(text) corpus un conjunto largo y estructurado de textos para el análisis.

document cada unidad del corpus

types para nuestro propósito, una única palabra

tokens cualquier palabra – conteo de tokens equivale al total de palabras

e.g. `A corpus is a set of documents.| This is the second document in the corpus.|`

es un corpus de dos documentos, donde cada documento es una oración. El primer documento tiene 6 types y 7 tokens. El segundo tiene 7 types y 8 tokens (por ahora ignoremos puntuación)

Conceptos básicos

stems palabras con sufijos removidos (usando un conjunto de reglas)

lemmas palabras en su forma canónica (la forma base de la palabra que tiene el mismo significado incluso cuando se le adiciona sufijos o prefijos)

word	win	winning	wins	won	winner
stem	win	win	win	won	winner
lemma	win	win	win	win	win

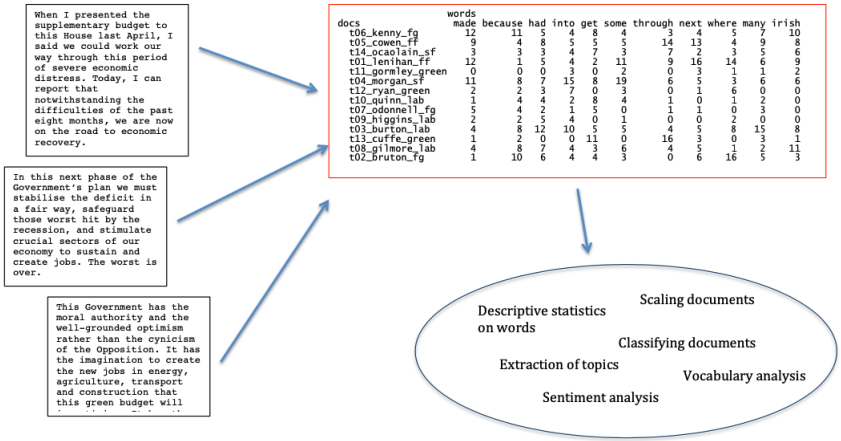
"key" words palabras seleccionadas por sus características especiales, significado, o sus tasas de ocurrencia

stop words palabras que son objeto de exclusión de cualquier análisis de texto.

Proceso fundamental de QTA

1. Selección de textos: definir el **corpus**
2. Conversión de textos a un formato electrónico común
3. Definir los documentos: decidir cuál va a ser la unidad de análisis documental
4. Definir los atributos. Esto puede tomar una variedad de formas, incluyendo **tokens**, clases de equivalencias de tokens (diccionarios), frases seleccionadas, características del lenguaje,...
5. Conversión de atributos textuales a una matriz cuantitativa (DFM)
6. Procedimiento estadístico o cuantitativo para extraer información de la matriz cuantitativa
7. Resumen e interpretación de los resultados cuantitativos

Proceso fundamental de QTA



Supuestos

1. El texto representa una implicación observable de alguna característica intrínseca de interés. Ej: un atributo del autor, un sentimiento o emoción, preponderancia de un tema (problema) político.
2. Los textos pueden ser representados a través de la extracción de sus características (*features*). La herramienta más común es la **bolsa de palabras** (*bag of words*).
3. Una **matriz documento-atributo** (*document-feature matrix*) puede ser analizada usando métodos cuantitativos para producir estimaciones de la característica de interés.

Bag-of-words approach

De palabras a números:

1. Preprocesamiento de texto:

"A corpus is a set of documents."

"This is the second document in the corpus."

Bag-of-words approach

De palabras a números:

1. Preprocesamiento de texto: cambiar a minúsculas,

"a corpus is a set of documents."

"this is the second document in the corpus."

Bag-of-words approach

De palabras a números:

1. Preprocesamiento de texto: cambiar a minúsculas, remover stopwords y puntuación,

"~~a~~ corpus ~~is~~ ~~a~~ set ~~of~~ documents~~.~~"

"~~this~~ ~~is~~ ~~the~~ second document ~~in~~ ~~the~~ corpus~~.~~"

Bag-of-words approach

De palabras a números:

1. Preprocesamiento de texto: cambiar a minúsculas, remover stopwords y puntuación, stem (dejar raíces),

   "corpus set documents"

   "second document corpus"

Bag-of-words approach

De palabras a números:

1. Preprocesamiento de texto: cambiar a minúsculas, remover stopwords y puntuación, stem (dejar raíces), tokenize en unigrams y bigrams (supuesto de bag-of-words)

   [corpus, set, document, corpus set, set document]

   [second, document, corpus, second document, document corpus]

Bag-of-words approach

De palabras a números:

1. Preprocesamiento de texto: cambiar a minúsculas, remover stopwords y puntuación, stem (dejar raíces), tokenize en unigrams y bigrams (supuesto de bag-of-words)

   [corpus, set, document, corpus set, set document]

   [second, document, corpus, second document, document corpus]

2. Document-feature matrix:
   - W: matriz de N documentos por M únicos n-gramas
   - w_{im} = número de veces que el m-ésimo n-grama aparece en el i-ésimo documenot.

	corpus	set	document	corpus set	⋮	M n-grams
Document 1	1	1	1	1	...	
Document 2	1	0	1	0	...	
...						
Document n	0	1	1	0	...	

Bag-of-words approach

- QTA frecuentemente desestima la gramática y el orden de las palabras y usa frecuencia de palabras como atributo.
- ¿Cuáles son las principales ventajas y limitaciones de este supuesto?

Frecuencia de palabras y sus propiedades

- El contexto es a menudo no es informativo: el uso de palabras individuales tiende a estar asociado con un grado particular de afecto, posición, ... sin importar mucho el contexto del uso de la palabra
- Las palabras solas tienden a ser más informativas ya que la ocurrencia de múltiples palabras (n-gramas) es raro.
- Algunos acercamientos se centran en la ocurrencia de una palabras como una variable binaria independiente de la frecuencia
- Otros acercamientos usan frecuencias: poisson, multinomial, entre otras distribuciones relacionadas

Estrategias de selección de atributos

¿Cómo elegir qué atributos incluir?

- **¿Todos?**

Estrategias de selección de atributos

¿Cómo elegir qué atributos incluir?

- **¿Todos?**:Computacionalmente ineficiente y las palabras raras generalmente no resultan informativas
- Recortar la DFM
- **Frecuencia de documento**: en cuántos documentos aparece el término
- **Frecuencia de término**: cuántas veces aparece el término en el corpus
- **Desestimación deliberada**: exclusión de palabras porque representan conectores lingüísticos de contenido no sustancial
- **Selección deliberada**: usar un diccionario de palabras o frases
- **Declaración de clases de equivalencia**: sinónimos no excluyentes (tesauro o *thesaurus*)

Stemming words

Lemmatization proceso algorítmico que se encarga de convertir palabras a lemmas.

stemming el proceso de reducir palabras compuestas a su raíz. Diferente a *lemmatization* en que stemmers opera en palabras sencillas sin conocimiento del contexto.

Ambas convierten las variantes morfológicas en términos raíz.

Ejemplo: `produc` a partir de
`producción, productor, producir, produces, produjo`

¿Por qué? Reducir el espacio de atributos al colapsar diferentes palabras en raíces (e.g. "felizmente" y "felicidad" transmiten el mismo significado que "feliz")