Short introduction to unstructured patient records' text mining

From regular expressions to text classification

Toni Mikkola

Tampere University Hospital • Pirkanmaa Hospital District

Clinical Informatics team

- Part of Tampere University Hospital research services
- We offer data services to researchers
- Our services include:
  - Data collection from hospitals’ systems
  - Data analysis and visualization
  - Machine learning
  - Etc.

Tays

What is Text Mining?

- Text Mining is the process of examining large collections of documents to discover new information or help answer specific research questions
- Variety of methodologies and tools to process the text
- Natural language processing (NLP)

Tays

Unstructured patient records

- Free texts inside the healthcare system
- Contain many different kinds of documents
  - Specialized care clinical texts (e.g. oncology…)
  - Nursing notes, nursing narratives
  - Laboratory, Pathology, Imaging
    - ➢ requests, notes, statements
  - Other kinds of free texts, e.g. specification of diagnosis

Tays

Characteristics of clinical texts 1/2

- Different from standard published texts
- Primarily written for hospital internal use and for mnemonic reasons
- Written by many different professionals with different writing skills and styles
- Contain sensitive personal information

Tays

Characteristics of clinical texts 2/2

- Often written under time pressure
  - ➢ misspellings, incomplete sentences, non-standard acronyms and abbreviations
- Use a lot of negations
- Speculative expressions
- This brings a lot of challenges to the text mining of clinical texts

Tays

When is it relevant?

- Wanted information (variable)
  - ➢ Not registered as structured information in EHR
    - at all, not in all departments or not during all the follow-up time period
  - ➢ Not trustable that the structured registration has been done
    - lack of time or rules for registration change over time and departments
- Cohort is too big for manual reading, or the researcher can’t have access to clinical texts

Tays

Regular Expressions (regex) 1/2

- Tool for finding expressions from text by patterns
- Textual syntax for representing patterns for matching text
- Need knowledge of how wanted information is registered in Electronic Health Records (EHR)
  - ➢ collaboration with local clinicians to find the right terms and stems for search

Tays

Regular Expressions (regex) 2/2

- Return positions and documents where the search terms appear for later reprocessing
- For extracting values
  - numerical, negative/positive, categorical
    - e.g. performance status (Karnofsky, Zubrod, WHO), gene mutations, TNM-stage
- For extracting sentences or text snippets around wanted terms for later analysis
- Highlighting terms in text for later analysis

Tays

Regex pros

- Easy to start, text doesn’t need any preprocessing
- Gets results fast
- Good for standard acronyms/abbreviations
- Good for ad hoc purposes

Tays

Regex cons

- Often needs manual work
  - manual classification of results
- Several iterations to find the best patterns
- Doesn’t give the full picture of text data
- Finds only what you search and everything else is hidden

Tays

Natural language processing, classical information retrieval 1/2

- For analyzing a collection of documents
- E.g. ”Bag-of-words” method
  - ➢ Each document is seen as the set of its terms
  - ➢ A term is not necessarily a word
    - Stemming - nursing -> nurs
    - Lemmatizing - nursing -> nurse
  - A lot of preprocessing and cleaning of documents
  - Models: boolean, vector space (tf-idf), statistical language models

Tays

Natural language processing, classical information retrieval 2/2

- Easy to query documents by search terms
  - Returns list of documents ranked by search terms
- Easy to find common terms used in documents
- Could also be used as a base for predicting models for document classification

Tays

Classical information retrieval - bag-of-words pros

- Simplified representation of text documents
- Quite agile to build, doesn’t necessarily need a lot of computation capacity
- Gives a full picture of the terms used in corpora
- Could be used to analyze and visualize corpora
- Allows to do queries to corpora
- Could be used to build document classification model
- Easy to compute similarities between documents

Tays

Classical information retrieval - bag-of-words cons

- Simplified model, losing information across the words
- Excludes grammar information and even the order of words
- Often excludes common words (stop words, e.g. “the”, “is”, “and”) and also punctuation marks from the model, to make it more simple

Tays

Deep learning -based language models

- A language model is basically a probability distribution over words or word sequences
- E.g. Google's Bidirectional Encoder Representations from Transformers (BERT)
- Many pretrained models in several languages are available as open source on the internet
  - So, no need to train the language model itself, you can download it

Tays

Text Classification models 1/2

- Classical IR models as tf-idf, or deep learning language models as BERT, could be used to build prediction models for classifying text documents
- E.g. smoking status (non smoker, ex-smoker or smoker), has the patient fallen (at the hospital, elsewhere or not at all)
- The model returns probability for given text to belong to each class

Tays

Text Classification models 2/2

- Supervised learning needs a labeled dataset to train the algorithm
  - ➢A lot of manual work
- Needs fine tuning and evaluation of the algorithm
- It’s a time-consuming process
- Sometimes, to get good accuracy of the model is difficult or almost impossible
- If the task is simple and repetitive, and has a lot of data, it is worth to try it

Tays

Conclusion

- Even if the searched information is not structurally registered and the cohort is too big, or going through all of the documents manually is impossible, there are still ways to collect variables for the researcher
- There are different methods to collect information from free text, but it often includes more inaccuracy than collecting information from structured data

Tays

Toni Mikkola
toni.mikkola@pshp.fi
Clinical informatics
Tays Research Services