

reviewAnalysis / README.md

jmarksk links, internal

yesterday

58 lines (31 loc) • 1.42 KB

Preview

Code

Blame

Raw



Drug reviews sentiment analysis

Background information

A hospital or insurance provider is interested in efficiently extracting numeric ratings from patients' written review. To this end we build a model using labelled, numerically, patient reviews.

Click [here](#) to access full notebook.

Data exploration

The data comes from Drugs.com and is accessed through UCI's website.

![Ratings distribution](./Images/Ratings distribution.png)

Click [here](#) to access data. Data was sourced by Kalummadi and Grer.

Ratings are not normally distributed. Counts are highest at the worst and best ratings.

- 160,000 samples
- 800 unique conditions
- 3400 unique drugs

Data understanding/preprocessing

- Tokenizing and creating the tf-idf matrix.

Data modeling [↗](#)

- Baseline Decision Tree model (TF-IDF)
- linear regression (TF-IDF)
- Word embedding models.

Results/conclusions [↗](#)



Recommendations/future work [↗](#)

- Deployment linear regression (TF-IDF) for 'rating extraction' from written review.
- Gather insights on how patients rate drugs.
 - "doctor, love, worse" etc.
- Combine the tf-idf and word embedding models.
- Use the "meta-data" as features. (i.e. the drug evaluated)

Repository structure [↗](#)

- [Notebook](#)
- README
- [Presentation](#)
- [data](#)