Summary:

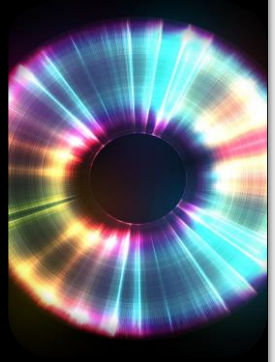
Clinical Text Pre-Processing for LDA (Latent Dirichlet Allocation)

Topic modeling using Latent Dirichlet Allocation (LDA) can be a powerful tool in identifying semantic groups in vast unstructured textual data like discharge notes. However, prior to modeling, text must be prepared, and often this step involves extensive pre-processing.

In this case, advanced pre-processing was applied to 1,000 discharge notes from the MIMIC-III data set. The final result can be used for

1. **Downstream modeling such as topic modeling with LDA.**
2. **Supervised and unsupervised text classification.**
3. **Similarity models such as TF – IDF.**

Repository on GitHub: https://github.com/olga12kz-DS/LDA-Pre-Processing



Text Pre-Processing Using:

- nltk
- spaCy
- sciSpaCy

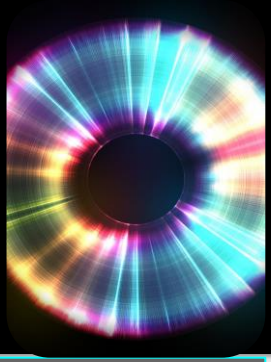
Extract “Present Illness History” field using User Defined Functions with **regex patterns and list comprehensions.**

Apply a **modified spaCy tokenizer** using the common English model and exclusions to 1) keep contractions, multi-word, and hyphenated words together; 2) keep % and /.

Identify top 110 abbreviations in the text and expand them.

Apply **lemmatization with POS tagging** for better precision; **lowercase, remove stopwords and punctuation**.

Special Rules for Keeping Integrity of Clinical Texts



Modifications to Standard Processing

1. # \d+% keep numbers and % signs together like 100%
   # \w+/\w+(?:/\w+)* keep slash-separated like f/v/c together
   # \w+(?:-\w+)+ keep hyphenated like Na-restricted together
   # \w+'\w+ keep contractions like don't together

2. List of top 100 abbreviations compiled from the text:

```
abbr_dict_manual = dict(sorted({
    "ASA": "aspirin",
    "AST": "aspartate aminotransferase",
    "BID": "twice a day",
    "BM": "bowel movement",…})
```

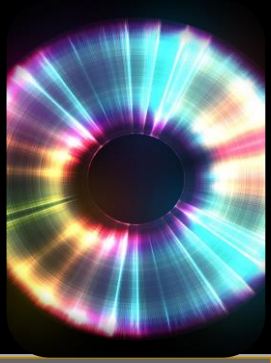
3. Lemmatization that splits multi-word and hyphenated words, lemmatized them, and re-joins them, applying POS tagging.

Examples of Output

From the note

Original	Output
17%	17%
f/v/c	fever/vomit/chill
Na-restricted	na-restrict
don’t	don’t’
emergency room	emergency room

From the Original Text to Final Results
Example of One Record



ORIGINAL TEXT

\n\n \nHistory of Present Illness:\n___ HCV cirrhosis c/b ascites, hiv on ART, h/o IVDU, COPD, \nbioplar, PTSD, presented from OSH ED with worsening abd \ndistension over past week. \nPt reports self-discontinuing lasix and spirnolactone ___ weeks \nago, because she feels like "they don't do anything" and that \nshe "doesn't want to put more chemicals in her." She does not \nfollow Na-restricted diets. In the past week, she notes that she \nhas been having worsening abd distension and discomfort. She \ndenies ___ edema, or SOB, or orthopnea. She denies f/c/n/v, d/c, \ndysuria. She had food poisoning a week ago from eating stale \ncake (n/v 20 min after food ingestion), which resolved the same \nday. She denies other recent illness or sick contacts. She notes \nthat she has been noticing gum bleeding while brushing her teeth \nin recent weeks. she denies easy bruising, melena, BRBPR, \nhemetesis, hemoptysis, or hematuria. \nBecause of her abd pain, she went to OSH ED and was transferred \nto ___ for further care...

FINAL PROCESSED TEXT

['hepatitis c virus', 'cirrhosis', 'c/b', 'ascites', 'hiv', 'antiretroviral therapy', 'h/o', 'intravenous drug use', 'chronic obstructive pulmonary disease', 'bioplar', 'post-traumatic stress disorder', 'present', 'outside hospital', 'emergency department', 'worsen', 'abd', 'distension', 'past', 'week', 'pt', 'report', 'self-discontinue', 'lasix', 'spirnolactone', 'week', 'ago', 'feel', 'like', 'anything', 'want', 'put', 'chemical', 'follow', 'na-restrict', 'diet', 'past', 'week', 'note', 'worsen', 'abd', 'distension', 'discomfort', 'denies', 'edema', 'shortness of breath', 'orthopnea', 'denies', 'fever , chill , nausea , and vomit', 'discharge or discontinue', 'dysuria', 'food', 'poison', 'week', 'ago', 'eat', 'stale', 'cake', 'n/v', '20', 'min', 'food', 'ingestion', 'resolve', 'day', 'denies', 'recent', 'illness', 'sick', 'contact', 'note', 'notice', 'gum', 'bleeding', 'brushing', 'teeth', 'recent', 'week', 'denies', 'easy', 'bruising', 'melena', 'bright red blood per rectum', 'hemetesis', 'hemoptysis', 'hematuria', 'abd', 'pain', 'go', 'outside hospital', 'emergency department', 'transfer', 'far', 'care', 'per', 'emergency department', 'report', 'pt', 'brief', 'period', 'confusion', 'recall', 'ultrasound', 'bloodwork', 'osh', 'denies', 'recent', 'drug', 'use', 'alcohol', 'use', 'denies', 'feel', 'confuse', 'report', 'forgetful', 'time', 'emergency department', 'initial', 'vitals', '98', '4', '70', '106/63', '16',