Generalni framework za semantičku analizu i klasterovanje Yelp recenzija

Aleksandar Ćurković

Matematički fakultet u Beogradu

27. septembar 2018.

Sadržaj

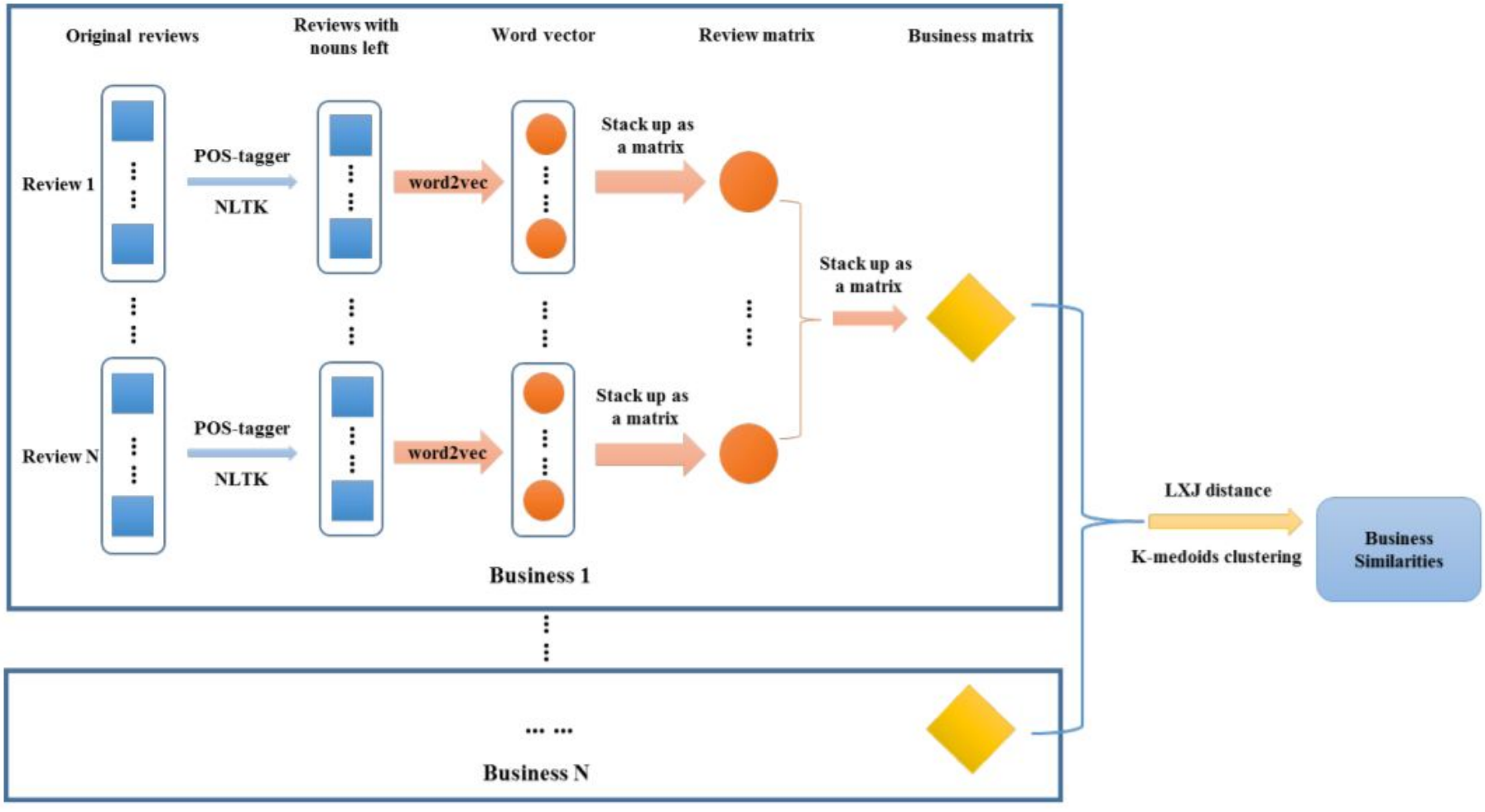
Uvod

- Automatska ekstrakcija korisnih informacija iz *Yelp* recenzija mogla bi biti veoma korisna kako za biznise tako i za korisnike
- *Word2Vec*, implementacija neuronskih mreža, pokazala je sposobnost da kvantifikuje semantičku sličnost izmedju reči
- Kao što samo ime sugeriše, svaka reč može se predstaviti kao vektor realnih brojeva

Metodologija

- Framework inkorporira *Word2Vec* kako bi reči iz recenzija konvertovao u vektore od kojih se kasnije formiraju matrice recenzija i biznisa
- Sa odgovarajuće definisanom distancom mogu se izvršiti analiza sličnosti i klasterovanje
- Primećeno je da Word2Vec nije u stanju da prepozna sentiment reči, pa se zbog toga koristi *Stanford-POS-Tagger* kako bi se iz recenzija izdvojile samo imenice

Metodologija



$$A=\begin{bmatrix} a_1 \\ \vdots \\ a_m \end{bmatrix} \quad B=\begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} \quad A \otimes B = \underset{i}{mean}\left\{ \max_j \left(\frac{a_i^T b_j}{\|a_i\|_2 \|b_j\|_2} \right) \right\} \quad LXJ(A,B) = \frac{A \otimes B + B \otimes A}{2}$$

Skupovi podataka

- Skupovi podataka podeljeni su u više datoteka, a za ovaj rad bile su dovoljne one koje se odnose na recenzije i biznise
- Svaka recenzija u sebi sadrži podatke kao što su o kom biznisu se radi, id korisnika koji je ostavio recenziju, broj zvezdica, i naravno, sam tekst recenzije
- Neki od bitnijih atributa biznisa su adresa, ime, ukupna recenzija i broj recenzija, kategorije primenjive na biznis kao i geografski podaci

Rezultati klasterovanja

- Korišćen je k-medoid algoritam
- Slučajno se bira 500 biznisa a algoritmu se kao broj klastera prosledjuje 20
- Uvidom u rezultate može se zaključiti da su instance u klasterima slične po *Yelp* tagu, odn. kategoriji biznisa kojim se bave

Korišćeni modeli i biblioteke

- Za Word2Vec korišćen je model *vsmlib* biblioteke koji je treniran nad podacima iz *Wikipedie*
  - Dimenzija vektorske reprezentacije reči u ovom modelu je 50
- Za tagovanje reči korišćen je *Stanford POS Tagger*
- Zbog velike količine podataka korišćena je *Dask* biblioteka, namenjena pre svega, skaliranju najčešće korišćenih paketa u *Python-u* kao što su *Numpy, Pandas* itd.

Korišćena literatura

- www.numpy.org
- pandas.pydata.org
- docs.dask.org/en/latest
- vsmlib.readthedocs.io/en/latest/index.html
- Bauckhage C. Numpy/scipy Recipes for Data Science: k-Medoids Clustering[R]. Technical Report, University of Bonn, 2015.