Izveštaj o izvršenju 6og zadatka

1. Odabir i priprema podataka:

- Odabrano je 20 skupova reči sa istim korenom na engleskom jeziku.

- Skupovi reči su definisani u Python listi `word_families_with_expected_stems`, gde je svaki skup predstavljen kao lista reči i očekivani stem.

- Reči i očekivani stemovi su sačuvani u fajl 'word_families_with_stems.txt'.

2. Učitavanje podataka:

- Reči i očekivani stemovi su učitani iz fajla 'word_families_with_stems.txt'.

3. Analiza stemmera:

- Implementirana je funkcija `evaluate_stemmer` koja evaluira performanse stemmera na odabranim skupovima reči.

- Analizirane su tri implementacije stemmera iz NLTK biblioteke: Porter, Lancaster i Snowball.

- Za svaki stemmer, izračunat je broj grešaka i identifikovane su problematične porodice reči.

4. Prikaz rezultata:

- Implementirana je funkcija `display_results` za prikaz rezultata evaluacije svakog stemmera.

- Prikazani su detalji o greškama i problematičnim porodicama reči za svaki stemmer.

5. Određivanje najboljeg stemmera:

- Upoređen je broj grešaka za svaki stemmer.

- Lancaster stemmer je identifikovan kao najbolji sa 17 grešaka, u poređenju sa 19 grešaka za Porter i Snowball stemmere.

6. Obrada dataset-a:

- Odabran je dataset 'email_classification.csv' koji sadrži email poruke i njihove oznake (spam ili ham) – vrlo slican datasetu sa vezbi.

- Implementirana je funkcija `preprocess_text` za obradu teksta, koja uključuje:

a) Tokenizaciju rečenica

b) Tokenizaciju reči

c) Uklanjanje stop reči

d) Primenu Lancaster stemmera (najbolji prema prethodnoj analizi)

7. Generisanje novog dataset-a:

- Učitan je originalni dataset.

- Primenjena je funkcija `preprocess_text` na svaku email poruku.

- Obrađeni podaci su sačuvani u novi CSV fajl 'processed_email_classification.csv'.

8. Verifikacija rezultata:

- Prikazani su originalni i obrađeni dataset-ovi korišćenjem pandas biblioteke, što omogućava vizuelnu potvrdu da je obrada uspešno izvršena.

Zaključak:

Zadatak je uspešno izvršen. Analizirana su tri stemmera, odabran je najbolji (Lancaster), i primenjen je na realnom dataset-u email poruka. Generisan je novi dataset sa obrađenim tekstom, koji uključuje tokenizaciju, uklanjanje stop reči i stemovanje. Ovaj obrađeni dataset može se dalje koristiti za zadatke klasifikacije ili druge analize teksta.