Mercari – price suggestion challenge

Bojana Đerić, Ivan Emanuel Pavlov,
Tatjana Ramljak, Lara Rajković

25. lipnja 2019.

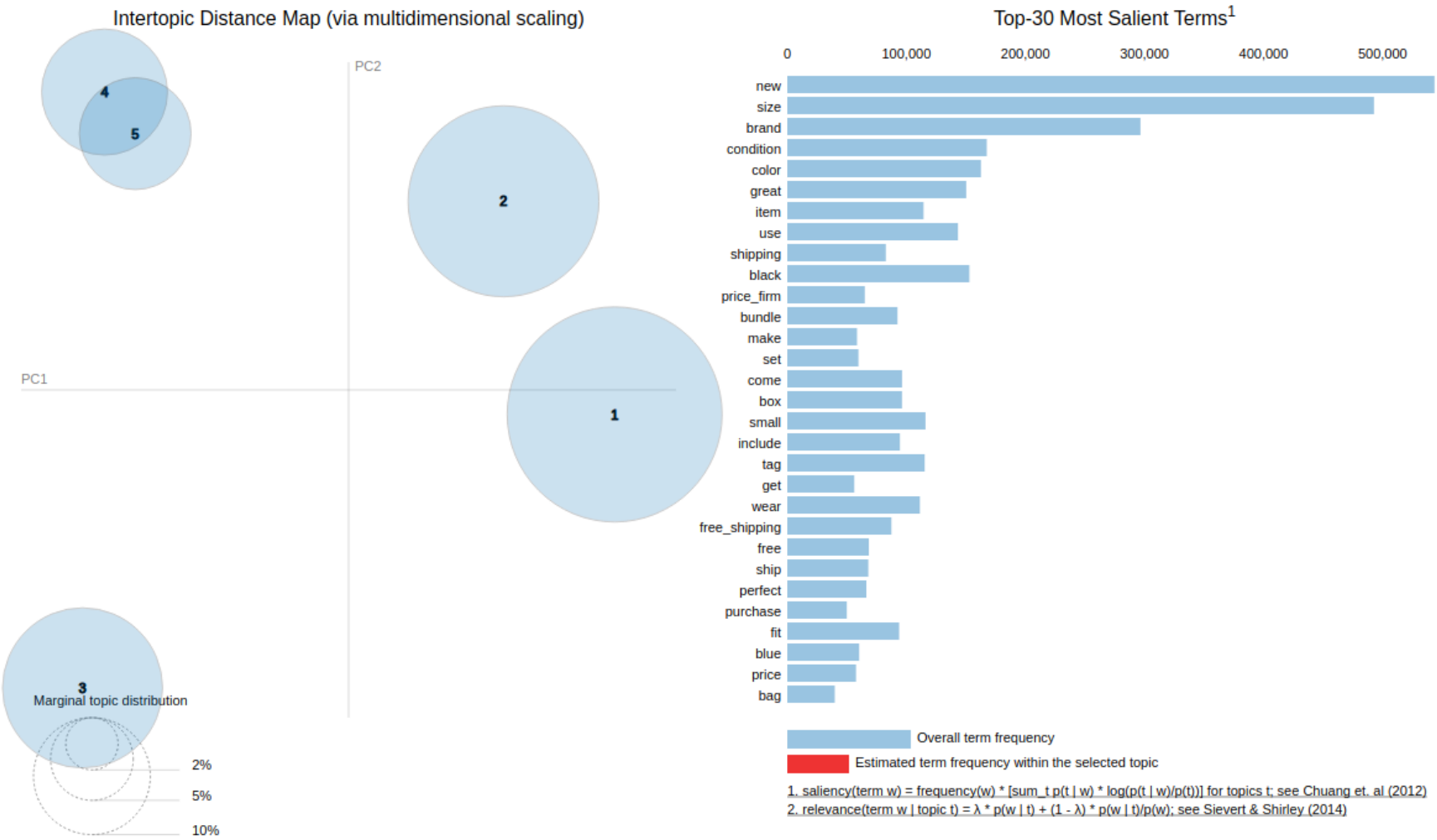
Uvodni opis problema

- Mercari – japanska aplikacija za prodaju i kupnju proizvoda.
- Natjecanje na Kaggleu
- Cilj: napraviti dobru predikciju cijene proizvoda
- Skup podataka za učenje sadrži 1482535 redova
- Značajke: `train_id`, `name`, `item_condition_id`, `category_name`, `brand_name`, `shipping`, `item_description`, ciljna varijabla `price`

Pretprocesiranje i *LDA*

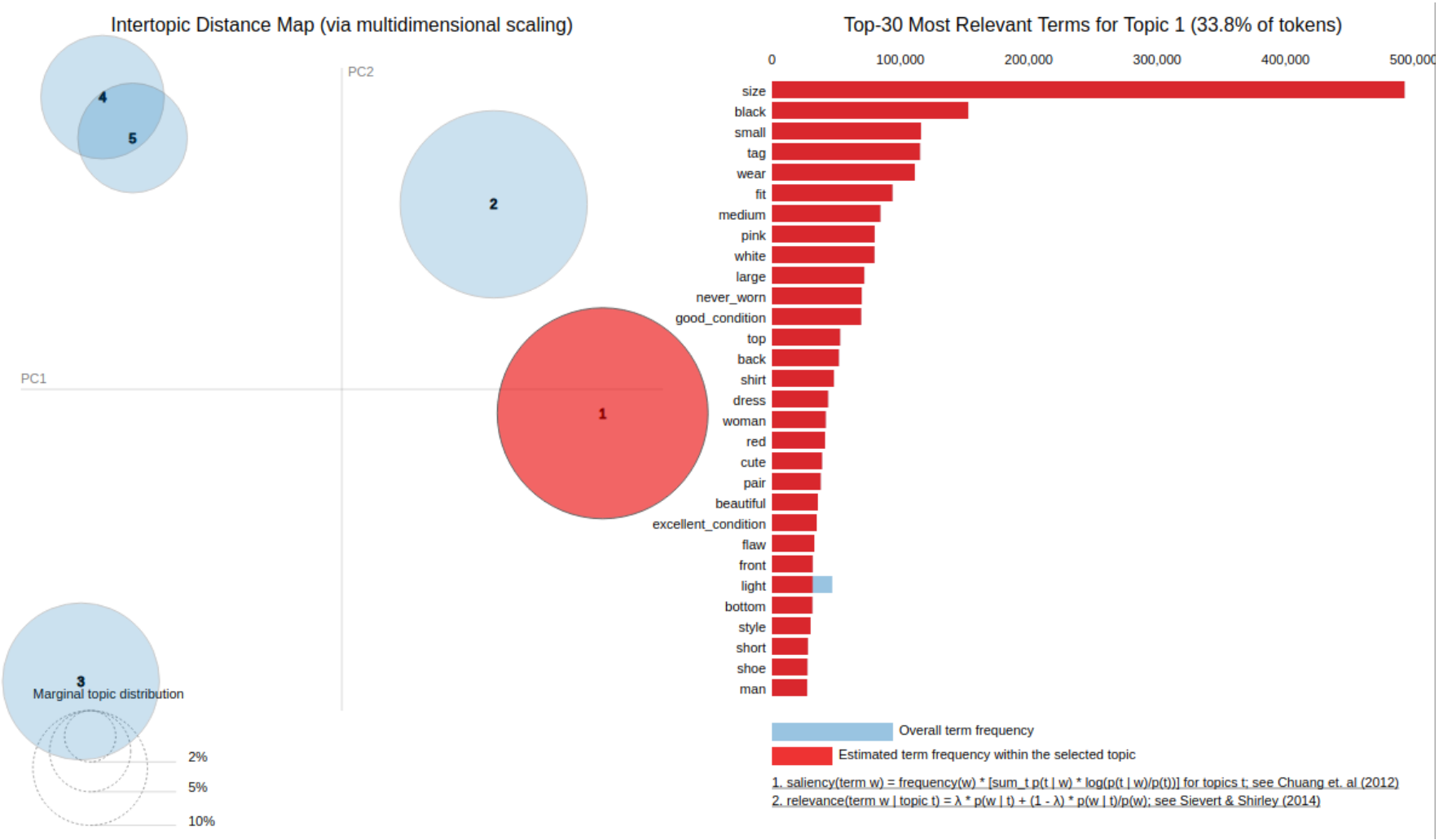
- `item_description` smo tokenizirali, izbacili *stop words*, stvorili bigrame i trigrame te lematizirali pomoću paketa *spacy* i *gensim*
- *Latent Dirichlet Allocation (LDA) – topic model* istreniran na 20, potom 5 tema
- dobivene distribucije tema u dokumentima – 5 novih značajki

LDA topic model



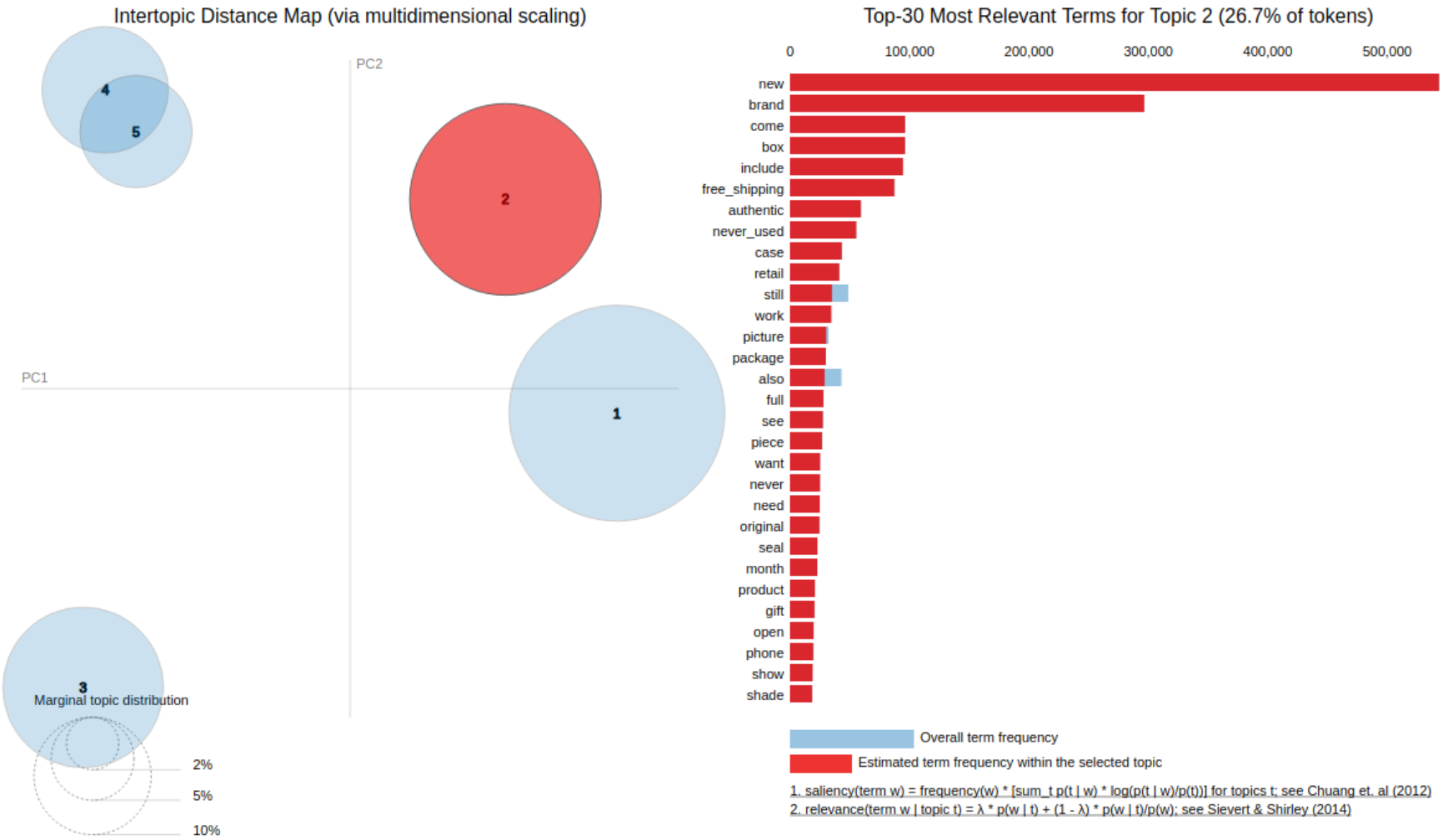
Slika: teme i frekvencije riječi

LDA topic model



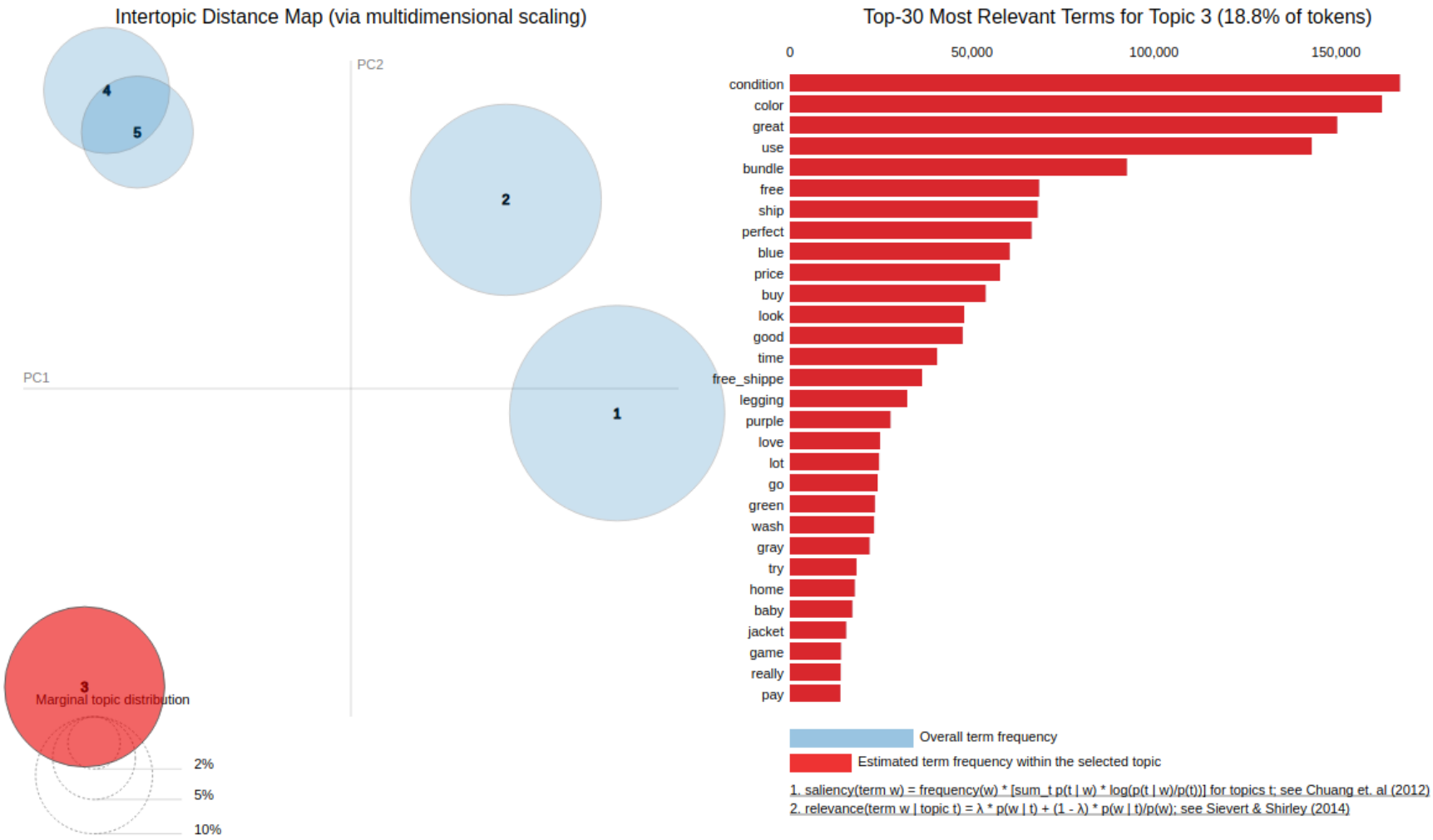
Slika: frekvencije riječi teme 1

LDA topic model



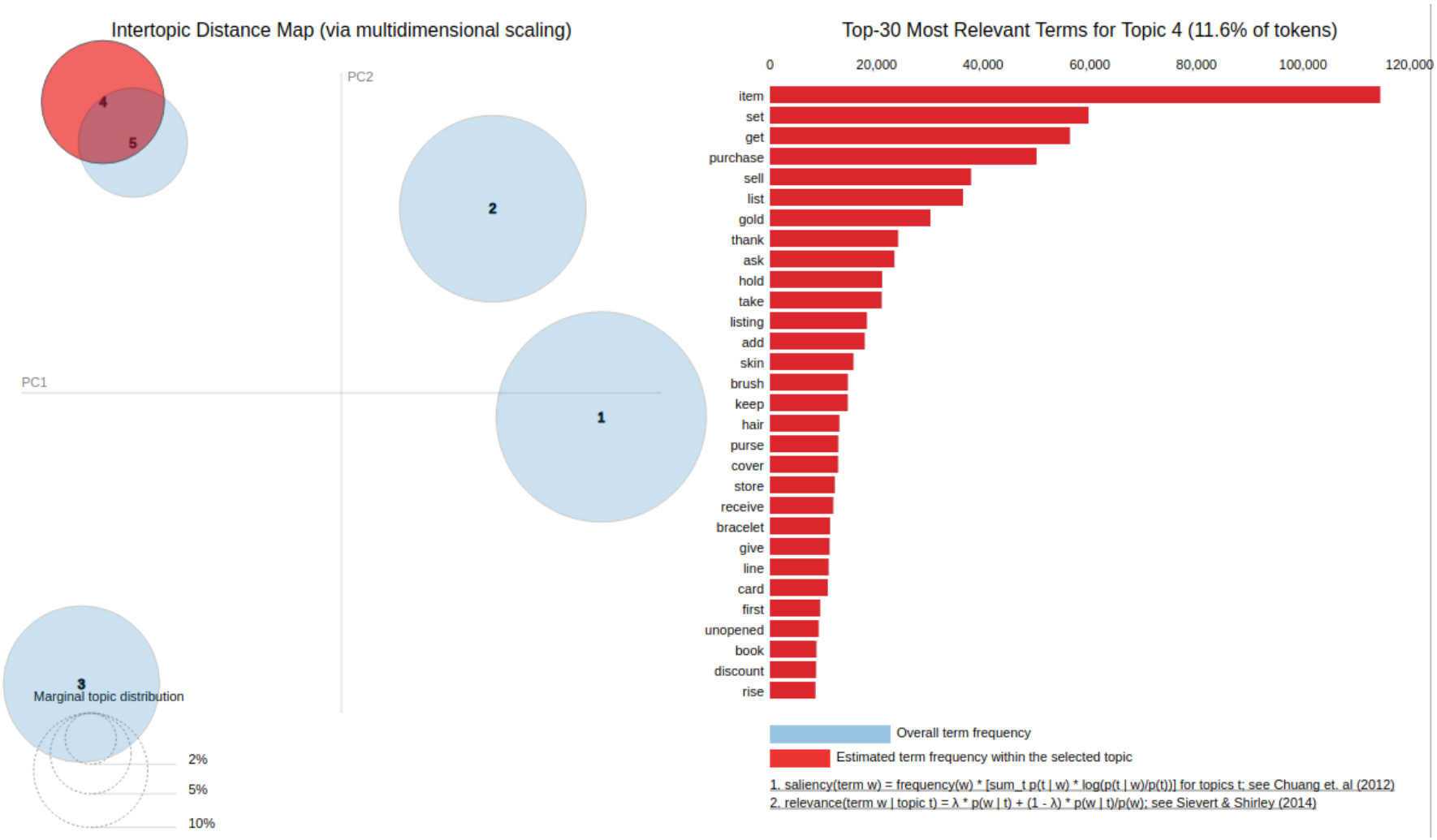
Slika: frekvencije riječi teme 2

LDA topic model



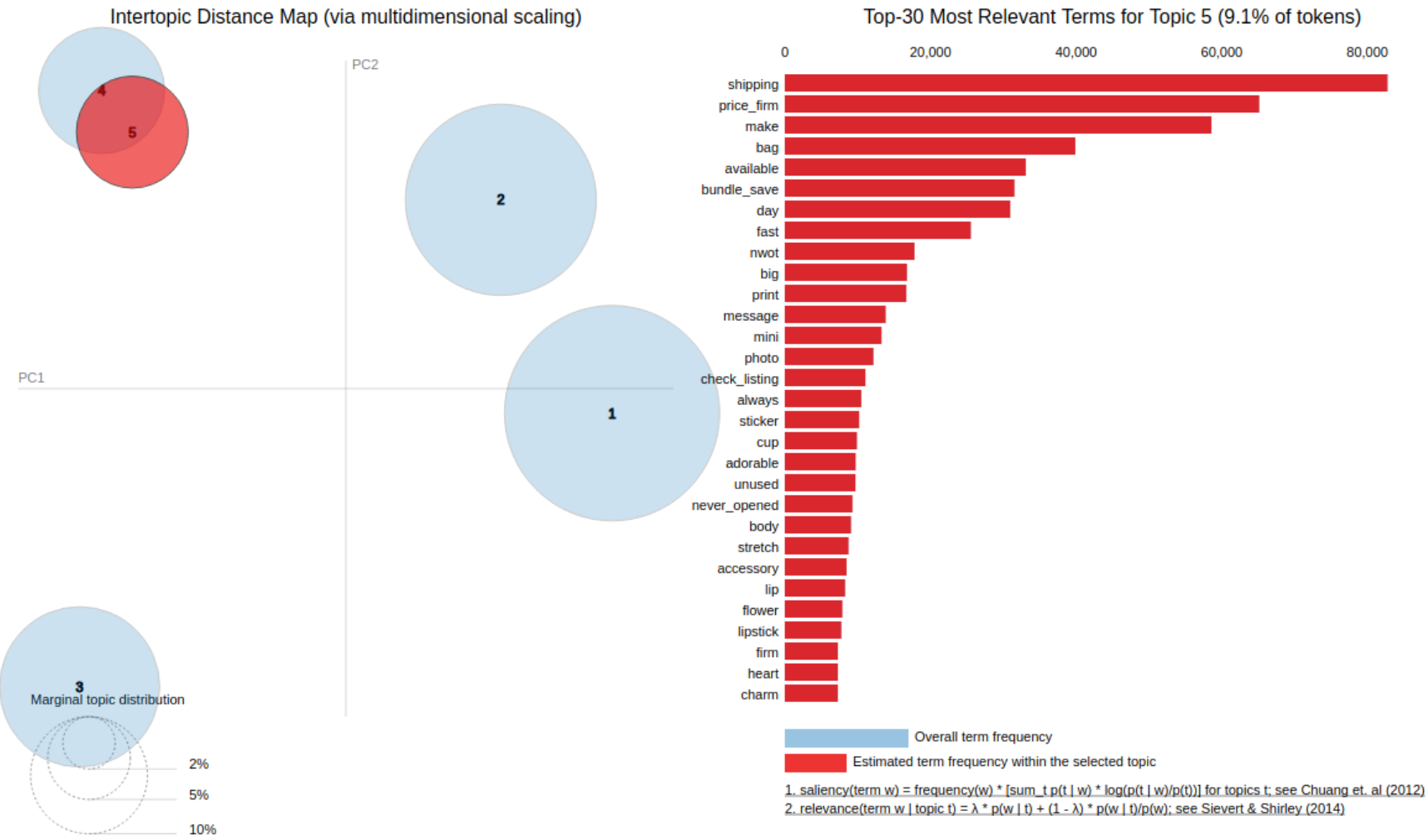
Slika: frekvencije riječi teme 3

LDA topic model



Slika: frekvencije riječi teme 4

LDA topic model



Slika: frekvencije riječi teme 5

Ridge regresija i *LightGBM* #1

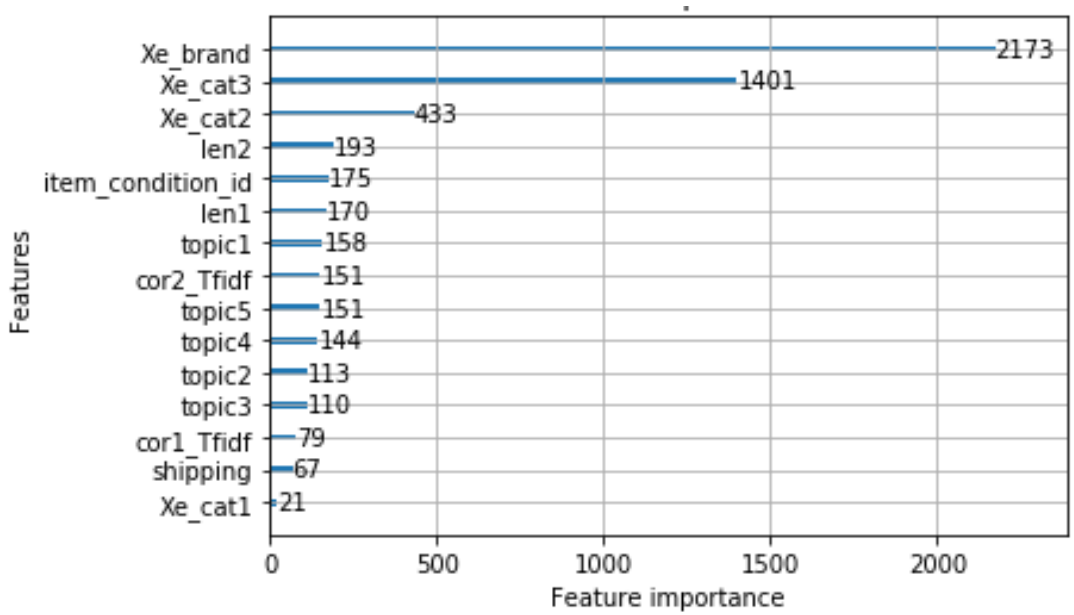
Pretprocesiranje:

- nedostajuće vrijednosti
- `item_condition_id` i `shipping` – *dummy* varijable
- `category_name` podijeljen u tri kategorije i transformiran pomoću *CountVectorizera*
- `brand_name` pomoću *LabelBinarizera*
- `item_description` pomoću *TfidfVectorizera*
- `name` pomoću *TfidfVectorizera* za *Ridge* i *CountVectorizera* za *LGBM*

Dobivena rijetka matrica dimenzije 1482535×161875

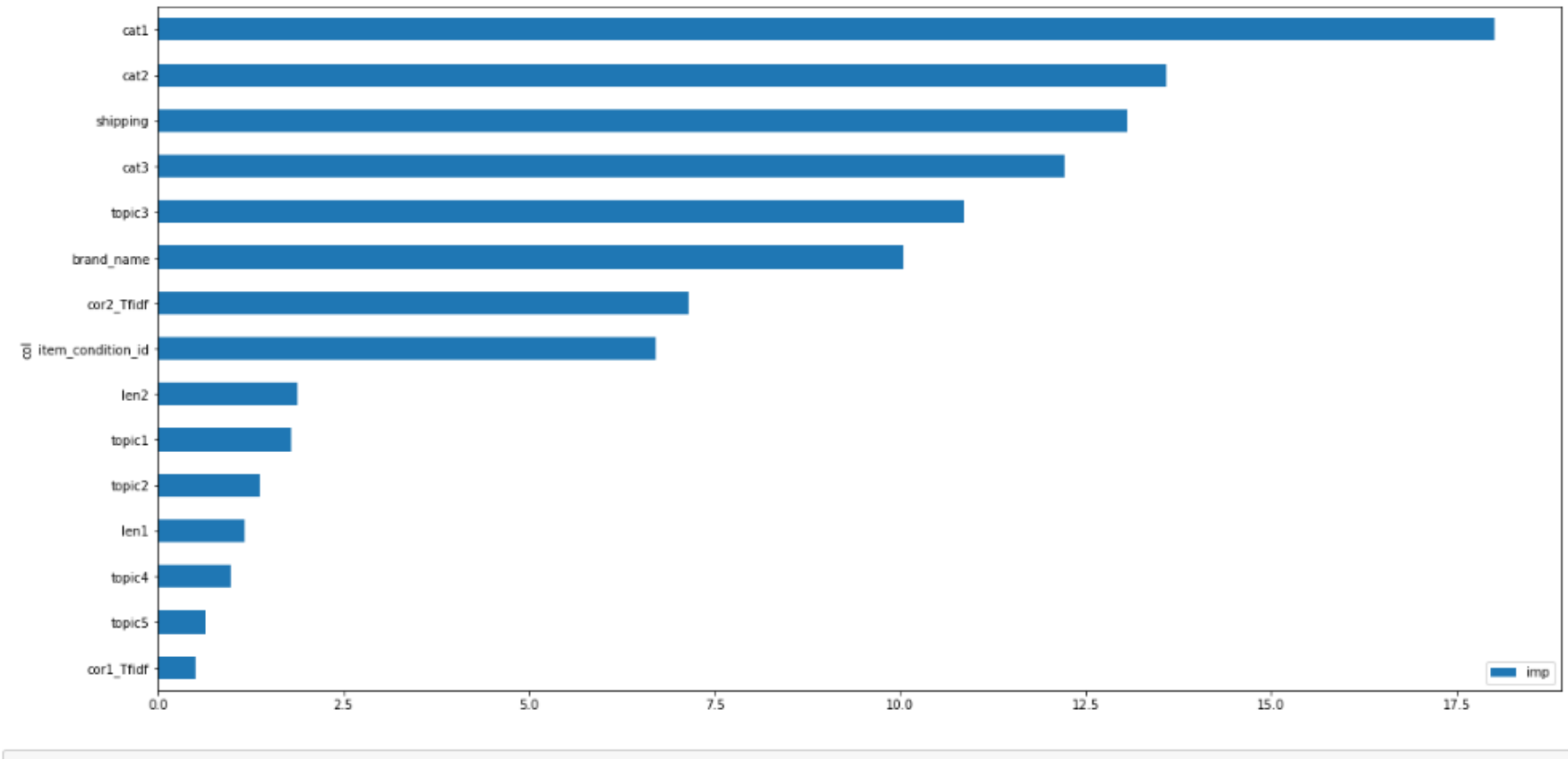
LightGBM #2 i CatBoost

- kategorijske varijable bez obrade
- dodane distribucije tema, srednje vrijednosti *tf-idf* težina i duljine stringova od `name` i `item_description`
- optimizacija parametara



Slika: *feature importance LGBMa*

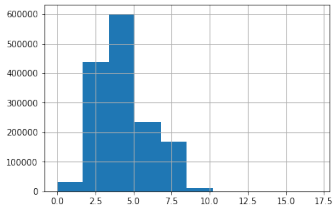
LightGBM #2 i CatBoost



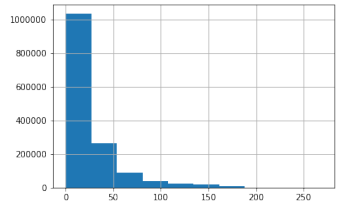
Slika: *feature importance CatBoosta*

Neuronske mreže - pretprocesiranje

- Nedostajuće vrijednosti
- Obrada kategorijskih varijabli `category_name` i `brand_name` → *LabelEncoder* iz *sklearn*a.
- Obrada tekstualnih varijabli `name` i `item_description` → tokenizacija *Kerasovim tokenizerom*
- Sekvence jednake duljine → *pad_sequences* → `name` ograničavamo na 10, `item_description` na 75.



Slika: Distribucija broja riječi u `name`



Slika: Distribucija broja riječi u `item_description`

Neuronske mreže - treniranje

- Korišteni su embedding layeri u Kerasu $\rightarrow$ variranje dimenzije embeddinga
- Kako su `name` i `item_description` tekstualne varijable *RNN layer*i su dobar kandidat za odabir aritekture.
- GRU tip RNN daje puno bolje rezultate od LSTM - 20 ćelija za `item_description` i 10 za *name*
- `Batch Normalization` i `Gradient Clipping` - nije pomoglo
- Isprobane su i *CNN*, ali nisu imale tako dobre rezultate
- Korišten *dropout* 0.1 (isprobani i drugi naravno)
- Povećanje broja epoha i smanjenje *batch size*
- Dodavanje *dense layera* s aktivacijskom funkcijom *RELU* (bolji rezultati) - isprobavali s raznim brojem neurona i broja layera
- Izlazni *layer* je jedan *dense layer* s linearnom aktivacijskom funkcijom $\rightarrow$ predviđamo kontinuiranu vrijednost

Ansambl

- $y_p = \sum_{i=1}^{5} w_i y_i$ pri čemu su w_i težine takve da je $\sum_{i=1}^{5} w_i = 1$, y_i su predikcije pojedinih modela i y_p je konačna predikcija.
- Težine su dobivene tako da greška predikcija bude što manja:
  - $w_1 = 0.021$ *LightGBM* s kategorijama
  - $w_2 = 0.379$ *LightGBM* s matricom
  - $w_3 = 0.068$ *Ridge* regresija
  - $w_4 = 0.016$ *CatBoost*
  - $w_5 = 0.516$ Neuronska mreža

Rezultati

- Originalni *training* skup podijeljen na novi *training* i validacijski skup u omjeru 95 : 5
- Računali smo *RMSLE* na validacijskom skupu
- $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(p_i+1)-log(a_i+1))^2}$
  pri čemu su
  n - broj podataka u skupu,
  p_i - predviđene cijene,
  a_i - stvarne cijene.

Rezultati

- *LightGBM* treniran na matrici daje rezultat 0.46105.
- *Ridge* regresija trenirana na matrici: 0.47196
- *CatBoost*: 0.52900
- *LightGBM* s kategorijama: 0.53438
- *RNN*: 0.44817 i 0.44212
- *CNN*: 0.45740
- **ansambl**: **0.43254**

Što smo naučili?

Hvala na pažnji!