

Facebook vestlusandmete analüüs ja mind imiteerivad chatbot'id



TARTU ÜLIKOOL
arvutiteaduse instituut

Andreas Vija
Informaatika 1. kursus

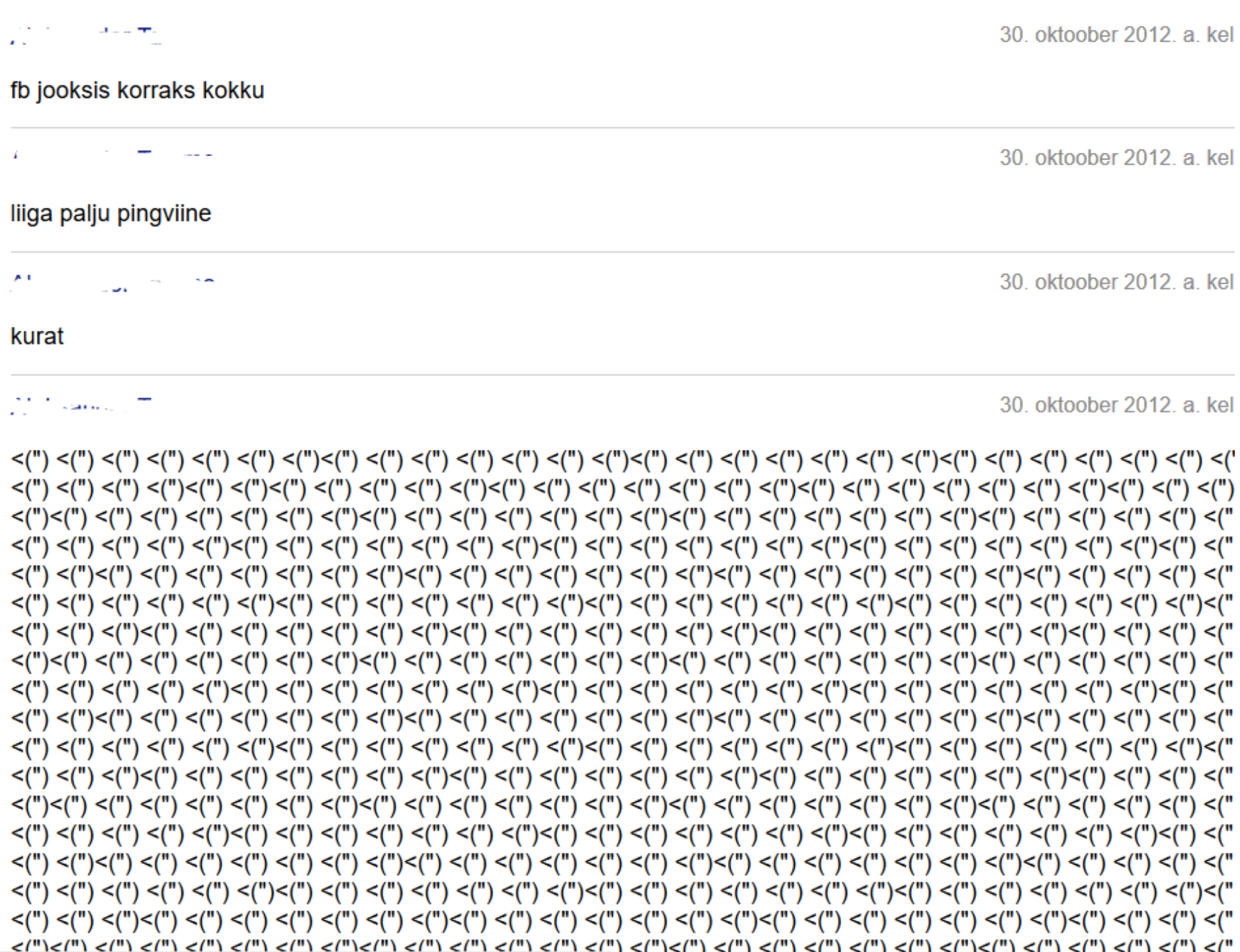


Sissejuhatus

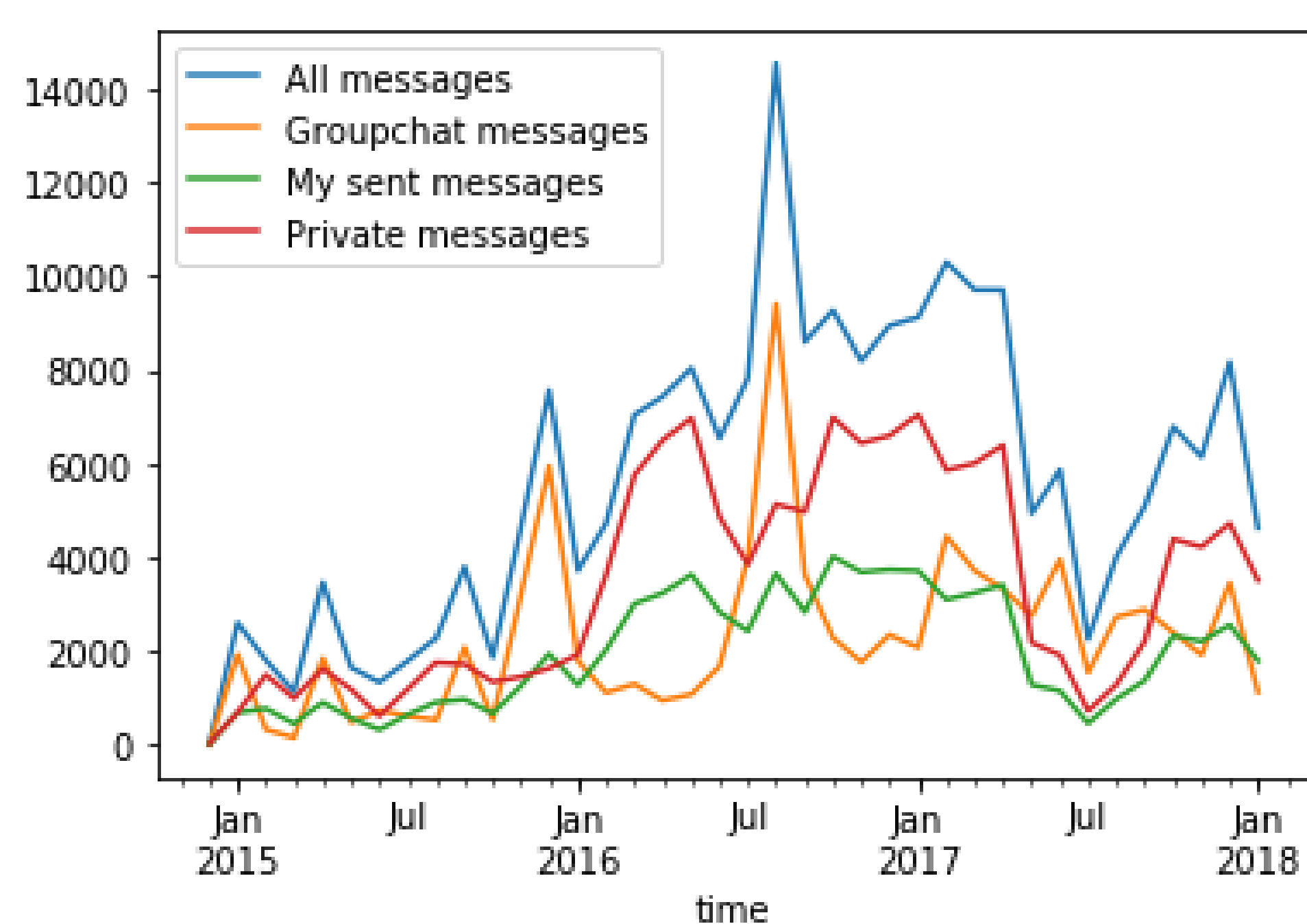
Facebook on juba tükk aega pakkunud kasutajatele võimalust tõmmata alla andmeid, mida nad on Facebooki sisestanud. Kuna need andmed on masintöödeldavas HTML formaadis, siis on võimalik nendega teha palju erinevaid asju. Enda vestlusandmete analüüs näiteks võib anda põnevat teavet enda käitumise ning selle muutumise kohta. Lisaks mõtlesin, et oleks väga huvitav panna chatbot mind jäljendama.

Suuremad tehnoloogiad

- **Python**, mugav ja kiiresti kirjutatav programmeerimiskeel, milles on võimsad teegid **pandas**, **matplotlib**, ja **numpy**.
- **gensim Doc2Vec**, Masinõppel põhinev tekstide sarnasuste ja seoste leidmise tehnoloogia.
- **Chatterbot**, Python teek, mis aitab vestluste põhjal luua chatbot'e.



Joonis 1. Peavalu tekitas üks 160k tähemärki pikk pingviinide rodu. Ei pandas, Pythoni konsool, ega Messenger ise ei talunud nii suurt sõnumit



Joonis 2. Sõnumite arv eri allikatest aja jooksul

Analüüs

Kasutasin Pythonit, et HTML andmed viia mugavale .csv kujule, ning seejärel uurisin andmeid pandas teegi abil – puhastasin andmeid veidi ning proovisin leida võimalikult palju visualiseeringuid ning statistilisi võrdlusi, mis võiks anda uut informatsiooni.



Joonis 3. Minule kõige iseloomulikumat sõnad

Analüüsi tulemused

- Hakkasin Messengeri aktiivsemalt kasutama 2015 lõpus.
- Sõnade arv ühes minu sõnumis on pidevalt tõusnud, samas sõltub see arv väga vestlusest.
- Kasutan ":D" rohkem kui enamus inimesi, kellega vestlen.
- Mu vestlusaktiivsus kasvab hommikust ööni järjepidevalt.
- Teistest päevadest veidi enam vestlen ma pühapäevadel ja esmaspäevadel.
- Projekti tegemise ajal olid Facebooki andmetes teatud vestlustes suured augud (mis nüüdseks on parandatud).

```
you: kuidas sul läheb?  
doc2vecbot: "norm ma 2"  
doc2vecbot answered in 0.8s  
chatterbot: "täitsa normilt võiks siiski rohkem tunde päevas olla"  
chatterbot answered in 24.4s  
  
you: kui pikk päev siis olla võiks?  
doc2vecbot: "neljapäev tähendab haha täiega norm pilt"  
doc2vecbot answered in 0.6s  
chatterbot: "nii pikk oligi"  
chatterbot answered in 24.6s  
  
you: aga mis siis kui päev oleks hoopis lühem?  
doc2vecbot: "kindel et saatsid õige pildi"  
doc2vecbot answered in 0.6s  
chatterbot: "sinine mis"  
chatterbot answered in 25.1s  
  
you: mis sa täna õhtul teed?  
doc2vecbot: "kindel et saatsid õige pildi"  
doc2vecbot answered in 0.6s  
chatterbot: "hunnik kingitusi rahvusmuustrites"  
chatterbot answered in 24.3s
```

Joonis 4. Lõige vestlusest loodud chatbot'idega

Chatbot'id

Masinõppe mudelite treenimiseks grupeerisin kõik oma vestlused sisendväljund paarideks, mida tuli kokku umbes 30 tuhat ja koostas suurt tekstikogumi - korpuse. Tegin kaks chatbot'i erinevatele alustehnoloogiatele:

Doc2Vec: Sügav närvivõrk õppis korpuse põhjal sõnadevahelisi seoseid ning paigutas etteantud tekstid selle põhjal 100-dimensionaalsesse vektorruumi. Selle põhjal on võimalik arvutada tekstide sarnasusi. Kasutaja sisendile vastati sellega, mida ma vastasin inimesele, kes kirjutas mulle kasutaja sisendile kõige sarnasema teksti.

Chatterbot: Bot treeniti vestluse põhjal ära tundma, mis vastused sobivad millistele tekstidele. Kuna chatbot õpib tervete vestluste pealt, pidi selleks, et ta omandaks just minu karakteri, andma andmeid vaid sisendväljund paaride kaupa, mis ei ole selle teegi algselt mõeldud kasutusviis. Kuna see bot oli nii aeglane, kasutati vaid 1/5 andmetest.

	ChatterBot	Doc2Vec
Vastamine	Aeglane	Kiire
Treenimine	Aeglane	Kiire
Relevantsus	Hea	Halb
Enese kordamine	Harv	Sage

Tabel 1. Kahe erineva chatbot'i võrdlus

Mida paremini teha?

- Ma ei ole kõige jutukam inimene, seega rohkem andmeid võimaldaks tõsisemat tekstianalüüsi ning paremaid chatbot'e.
- Keerulisemad tekstianalüüsi meetodid.
- Suurem korpus kui ainult minu vestlused, näiteks mõni estnlitk korpus.
- Sofistikeeritum algoritm Doc2Vec kasutamiseks.
- Chatterbot ei sobi minu eesmärgiks hästi.

Kontaktinfo:

Andreas Vija
Informaatika bakalureuse 1. kursus
Arvutiteaduse instituut, loodus-ja täppisteaduste valdkond
andreasvija@hotmail.com
github.com/andreasvija
[linkedin.com/in/andreasvija/](https://www.linkedin.com/in/andreasvija/)

Projekti asukoht:

github.com/andreasvija/chat-analysis
github.com/andreasvija/chatbots-of-me