DataSentics

Towards *chatbotization* of POC on George chats

powered by databricks Microsoft Azure

DataSentics in a nutshell

How We Work

ML Center of Excellence in Prague & Brno

Machine learning 30+

Data engineering Cloud 25+

Partners of

databricks Microsoft aws

Key References

ČESKÁ spořitelna · MONETA · Allianz · VIG Vienna Insurance Group · Asahi · MALL · IKEA · dentsu aegis network · pepsi · socialbakers · Red Bull

Ad Picker

Customer Engagement 360°
Combining digital and offline behavioral data and machine learning to personalize the customer experience

Fraud / AML / Risk
Using advanced methods such as NLP and graph analytics to identify anomalies, etc.

Contact center
Using NLP for automatic routing of emails/request, topic detection, etc.

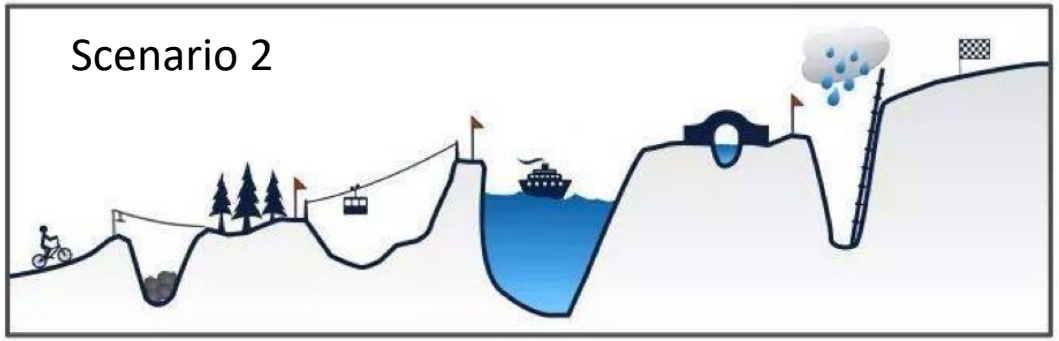
Quality inspection
Computer-vision-based AI solution for automatically analyzing the quality of placement of my products in a store shelf

SHEL INSPECTOR

*"Make data science and **machine learning** have a **real impact** on organizations across the world...*

*...bring to life **transparent production-level data science**.*

Building chatbot



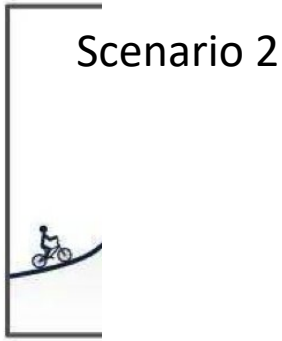
Building chatbot

Scenario 2: milestones along the way

Building chatbot

Scenario 2: milestones along the way

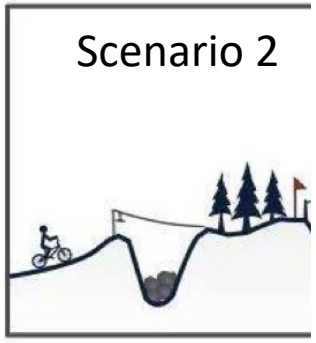
1. Parse chats (lemmas, relations, stop-words, …)



Building chatbot

Scenario 2: milestones along the way

1. Parse chats (lemmas, relations, stop-words, …)
2. Operational dashboard (number of chats, word frequencies, …)



Building chatbot

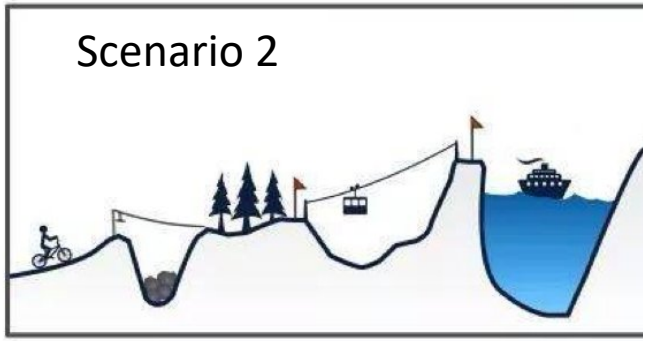
Scenario 2: milestones along the way

1. Parse chats (lemmas, relations, stop-words, …)
2. Operational dashboard (number of chats, word frequencies, …)
3. Alerting (significant increases/decreases)

Building chatbot

Scenario 2: milestones along the way

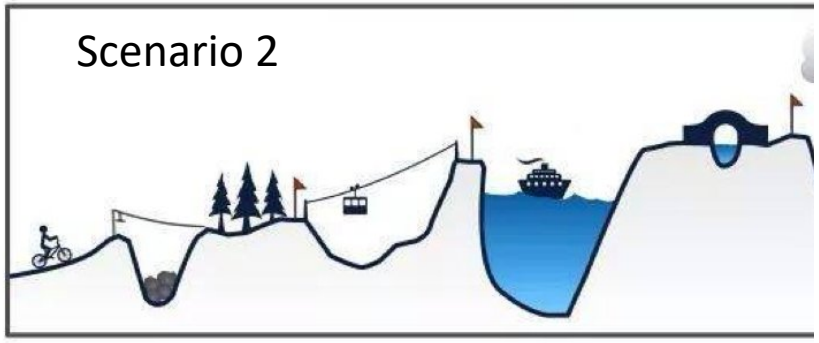
1. Parse chats (lemmas, relations, stop-words, …)
2. Operational dashboard (number of chats, word frequencies, …)
3. Alerting (significant increases/decreases)
4. Routing (classify chats into known topics)



Building chatbot

Scenario 2: milestones along the way

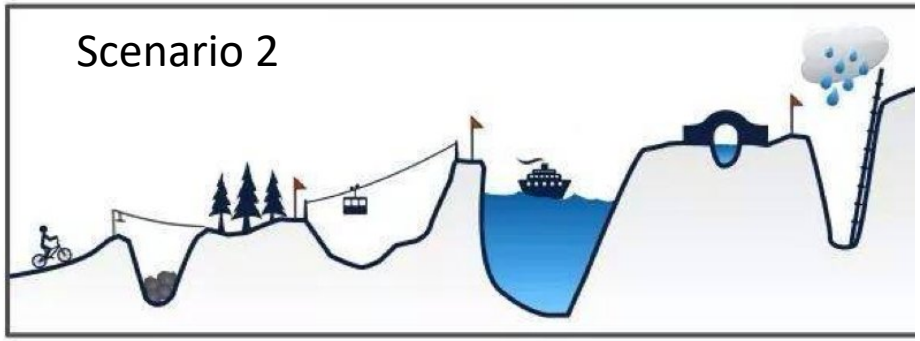
1. Parse chats (lemmas, relations, stop-words, …)
2. Operational dashboard (number of chats, word frequencies, …)
3. Alerting (significant increases/decreases)
4. Routing (classify chats into known topics)
5. Indentification of new topics



Building chatbot

Scenario 2: milestones along the way

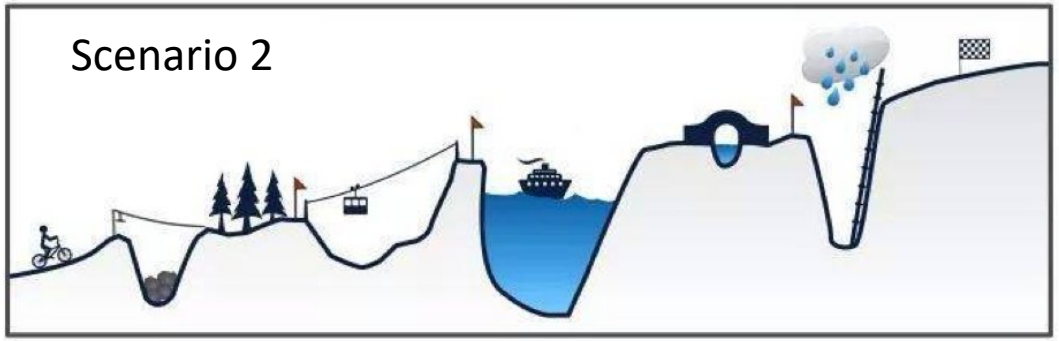
1. Parse chats (lemmas, relations, stop-words, …)
2. Operational dashboard (number of chats, word frequencies, …)
3. Alerting (significant increases/decreases)
4. Routing (classify chats into known topics)
5. Indentification of new topics
6. FAQ/QA



Building chatbot

Scenario 2: milestones along the way

1. Parse chats (lemmas, relations, stop-words, …)
2. Operational dashboard (number of chats, word frequencies, …)
3. Alerting (significant increases/decreases)
4. Routing (classify chats into known topics)
5. Indentification of new topics
6. FAQ/QA
7. …
8. …
9. …
10. Chatbot



Chats parsing

What we did:

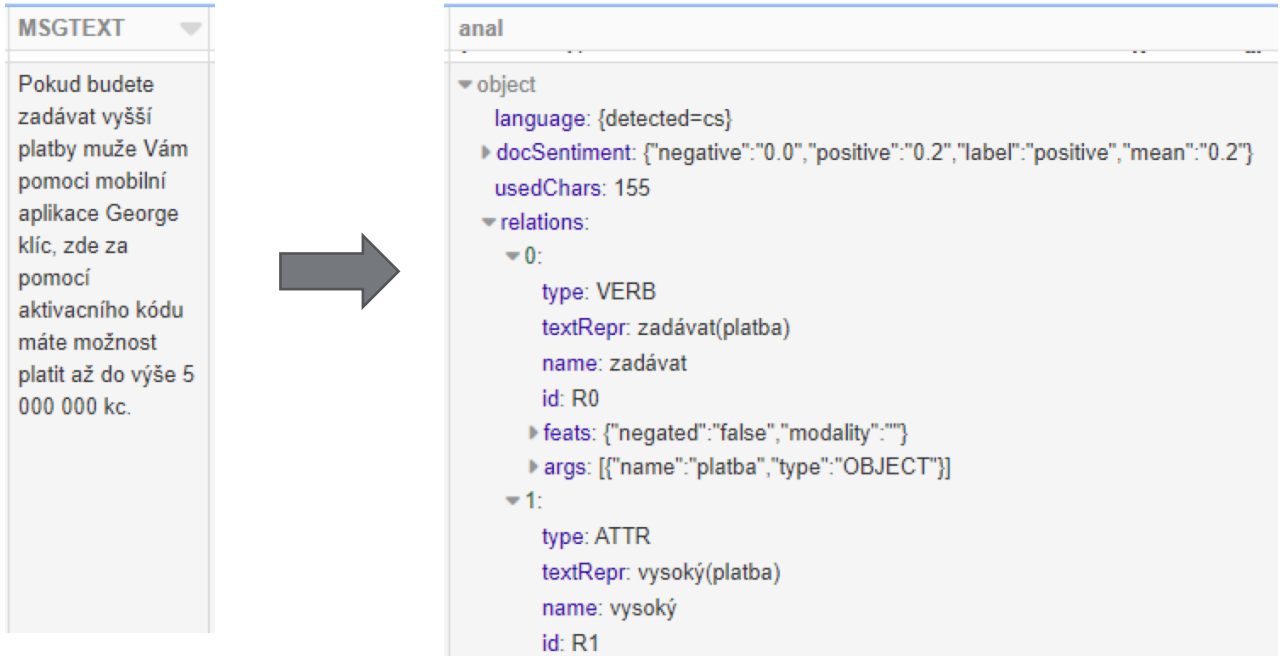
- *Parsing csv export*
- *Text preprocessing (IDs, Geneea lemmatization, stopwords)*
- *Operator/client identification*

Output:

- *Right structure for further analysis*
- *Relevant text representations*

Benefits:

- *Chats ready in Azure for other use cases*
- *Sales signals, prevention, …*



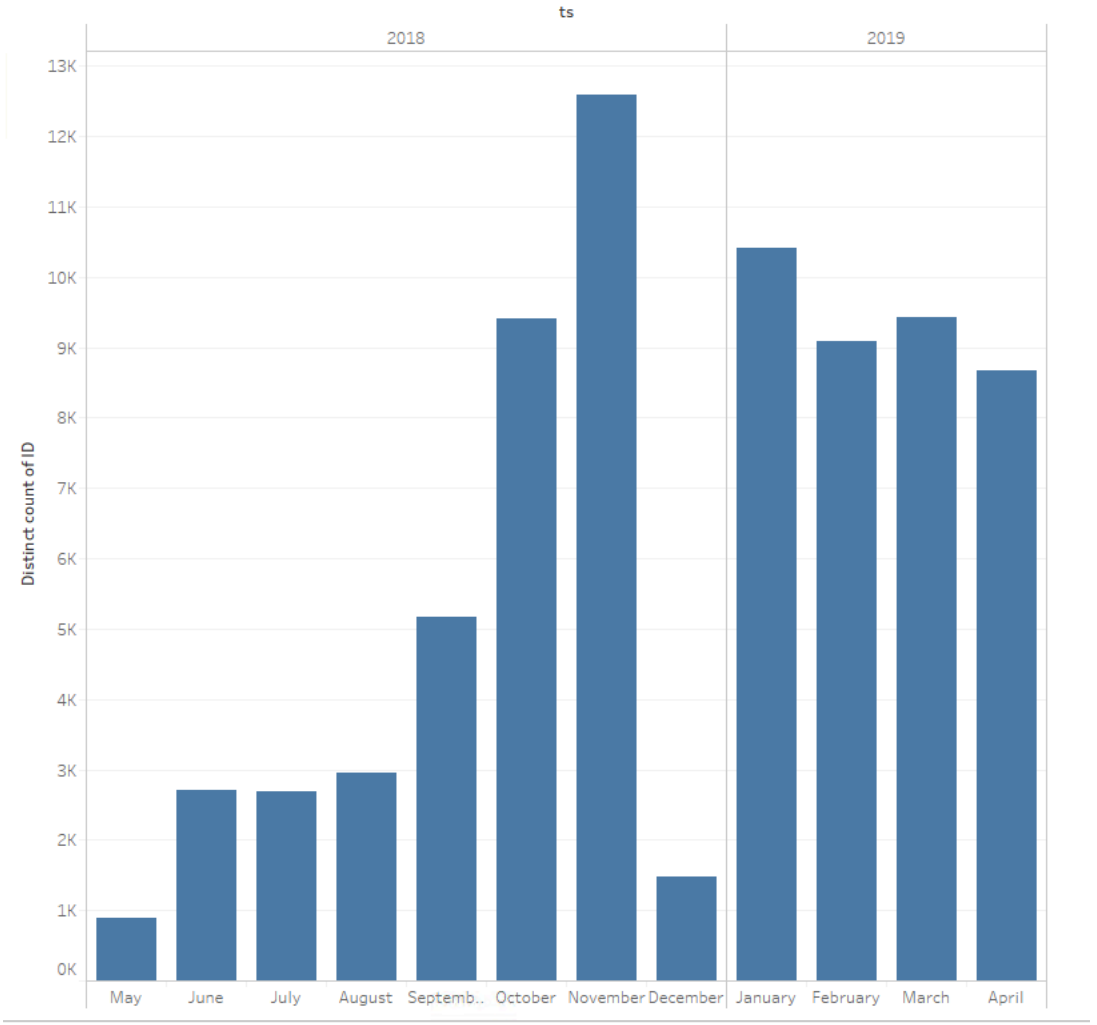
Operational dashboard

What we did:

- *Number of chats per specific periods (year, month, …) and dimensions (category, result, operator/client)*
- *Number of people in the chats (routing)*
- *Most frequent and significant relations per specific periods/dim.*

Benefits:

- *Overview whats going on in the chats*



Operational dashboard (relations per month, % of chats)

DataSentics

ts

Text Repr	2018 May	June	July	August	Septemb..	October	November	December	2019 Janua..	February	March	April
parent(bankovní produkt,..	0.4016	0.3933	0.3856	0.4252	0.3972	0.3704	0.3857	0.4114	0.2976	0.3554	0.3315	0.4373
parent(finanční instituce,..	0.0774	0.0725	0.0759	0.0996	0.0855	0.0777	0.0925	0.0952	0.1784	0.0916	0.0920	0.1099
parent(měna,CZK)	0.1081	0.1121	0.1304	0.1375	0.1408	0.1298	0.1399	0.1619	0.1286	0.1817	0.1919	0.2438
parent(bankovní produkt,..	0.1684	0.1671	0.1785	0.1861	0.1618	0.1551	0.1785	0.1814	0.1281	0.1990	0.1967	0.2557
clearingový(kód)				0.0003	0.0002	0.0007	0.0011	0.0035	0.1194	0.0014	0.0023	0.0020
ozubený(kolečko)	0.0182	0.0333	0.0338	0.0352	0.0545	0.0695	0.1042	0.0778	0.1188	0.1030	0.0792	0.0710
parent(finanční instituce,..	0.1160	0.1293	0.1255	0.1269	0.1067	0.0961	0.0983	0.1015	0.0905	0.0991	0.0989	0.1216
parent(banka,Česká Spoři..	0.1160	0.1293	0.1255	0.1269	0.1067	0.0961	0.0983	0.1015	0.0905	0.0991	0.0989	0.1216
uvedený(kód)		0.0004			0.0002	0.0004	0.0008	0.0014	0.0893	0.0008	0.0014	0.0012
vybraný(země)						0.0002	0.0007	0.0014	0.0865	0.0009	0.0012	0.0009
aktualizovat(stránka)		0.0191	0.0210	0.0243	0.0373	0.0498	0.0254	0.0278	0.0690	0.0218	0.0376	0.0184
parent(bankovní produkt,..	0.0262	0.0269	0.0203	0.0157	0.0186	0.0173	0.0139	0.0181	0.0670	0.0799	0.0533	0.0319
telefonní(číslo)	0.0785	0.0568	0.0695	0.0763	0.0672	0.0545	0.0519	0.0577	0.0649	0.0821	0.0678	0.0710
technický(podpora)	0.0239	0.0355	0.0560	0.0694	0.0807	0.0834	0.0681	0.0598	0.0593	0.0471	0.0385	0.0365
levý(sloupec)	0.0148	0.0183	0.0410	0.0434	0.0567	0.0568	0.0662	0.0549	0.0577	0.0651	0.0532	0.0527
žádat(platba)	0.0990	0.0501	0.0507	0.0554	0.0615	0.0629	0.0645	0.0695	0.0572	0.0475	0.0370	0.0473
parent(transakce,převod)	0.0478	0.0654	0.0661	0.0759	0.0682	0.0620	0.0660	0.0709	0.0571	0.0634	0.0910	0.0837
parent(bankovní produkt,..	0.0341	0.0396	0.0365	0.0397	0.0293	0.0255	0.0228	0.0264	0.0543	0.0629	0.0493	0.0514

Alerts (period significant key word)

What we did:

- *Key words (topic-like) extractions using TF-IDF (term frequency – inverse document frequency) per category*
- *Key words trends per specific periods*

Benefits:

- *Phrases for daily/weekly alerts – somethings wrong in the app*
- *Slack/Email/…*

CATEGORY_num	CATEGORY	tag_0	tag_1	tag_2	tag_3	tag_4	tag_5	tag_6	tag_7	tag_8	tag_9
01	Jan	vybraný(země)	clearingový(kód)	uvedený(kód)	zaplacený(úrok)	aktualizovat(stránka)	najít(potvrzení)	současný(stisknutí)	daňový(přiznání)	prosit(stránka)	děkovat(napsání)
02	Feb	zaplacený(úrok)	najít(potvrzení)	daňový(přiznání)	zvolit(produkt)	děkovat(napsání)	aktualizovat(stránka)	kliknout(produkt)	stáhnout(klíč)	hezký(dopoledne)	poslední(výpis)
03	Mar	zaplacený(úrok)	spustit(průvodce)	daňový(přiznání)	dokončit(přechod)	aktualizovat(stránka)	najít(potvrzení)	on-line(průvodce)	stáhnout(klíč)	děkovat(napsání)	služba,správa
04	Apr	služba,správa	zaplacený(úrok)	stáhnout(klíč)	aktualizovat(stránka)	děkovat(napsání)	modrý(lišta)	zvolit(produkt)	hezký(dopoledne)	kliknout(produkt)	potřebovat(kód)
05	May	prověřit(situace)	těšit	modrý(lišta)	pěkný(víkend)	daňový(přiznání)	příští(platba)	zvolit(adresář)	platný-not(datum)	prověřit(případ)	děkovat(chat)
06	Jun	modrý(lišta)	těšit	prověřit(situace)	aktualizovat(stránka)	služba,správa	kliknout(produkt)	pěkný(víkend)	vymazat(údaj)	platný-not(datum)	platit(platební
07	Jul	modrý(lišta)	aktualizovat(stránka)	prověřit(situace)	služba,správa	kliknout(produkt)	chytrý(telefon)	platit(platební	spustit(průvodce)	nastavit(oprávnění)	on-line(průvodce)
08	Aug	aktualizovat(stránka)	modrý(lišta)	děkovat(napsání)	nastavit(oprávnění)	chytrý(telefon)	prověřit(situace)	těšit	potřebovat(kód)	služba,správa	platit(platební
09	Sep	aktualizovat(stránka)	chytrý(telefon)	modrý(lišta)	hezký(dopoledne)	děkovat(napsání)	klávesový(zkratka)	stáhnout(klíč)	odpovědět(rád)	horní(panel)	zvolit(produkt)
10	Oct	aktualizovat(stránka)	hezký(dopoledne)	stáhnout(klíč)	prosit(stránka)	modrý(lišta)	odpovědět(rád)	klávesový(zkratka)	chytrý(telefon)	kliknout(produkt)	ozubený(kolo)
11	Nov	aktualizovat(stránka)	odpovědět(rád)	kliknout(produkt)	zvolit(produkt)	odpovědět(den)	stáhnout(klíč)	ozubený(kolo)	zvolit(adresář)	současný(stisknutí)	vygenerovat
12	Dec	aktualizovat(stránka)	služba,správa	současný(stisknutí)	clearingový(kód)	váš(produkt)	potřebovat(kód)	nastavit(oprávnění)	zvolit(produkt)	kliknout(kolečko)	kliknout(produkt)

Topics from voice chats: manual tagging of training set

Text analysis:

- *Reasons from voice chats to relations representation*
- *Search relations in chats (60%)*

Benefits:

- *Routing*
- *Explanation of chats*

nastavení limitů	co je to za limit 200 tis. , kde ho mohu změnit
	limity u karet - nemohu dohledat, kde nastavím limit pro kartu
	nemohu najít nastavení limitů

```
limity = [['nastavit(limit)'],
          ['parent(bankovní produkt,platební karta)'],
          ['najít(nastavení)'],
           ['změnit(limit)']] # nastaveni limitu
```

Supervised model

MSGTEXT
Pokud budete zadávat vyšší platby muže Vám pomoci mobilní aplikace George klíc, zde za pomocí aktivacního kódu máte možnost platit až do výše 5 000 000 kc.

NUMERIC REPRESENTATIONS

Standard classification models

(GLMs as benchmark, GBTs in production)

High Dimensional representation of words using word2vec

DataSentics

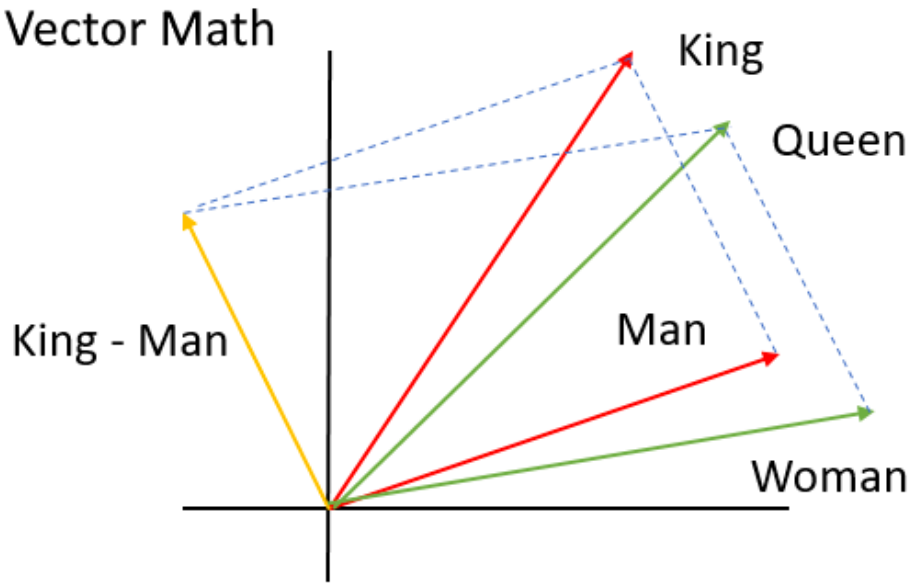
- *Can be trained on a very large unstructured text (no need for annotation).*
- *Very powerful in finding semantic and syntactical similarities*
- *Each word can be converted into vector (word embeddings) -> Enables mathematical operations on words (vectors):*

Vec(king) – Vec(man) + Vec(woman) = Vec(queen)

```
1 synonyms = model.findSynonyms("dům", 10)
2 display(synonyms.filter(synonyms.similarity>0.5))
```

▸ synonyms: pyspark.sql.dataframe.DataFrame = [word: string, similarity: double]

word	similarity
domek	0.6293026208877563
pozemek	0.6255326271057129
byt	0.5934215784072876
hotel	0.5879210829734802
objekt	0.5776727795600891
stallburg	0.5513975024223328
komplex	0.5492702722549438
areál	0.5435788631439209



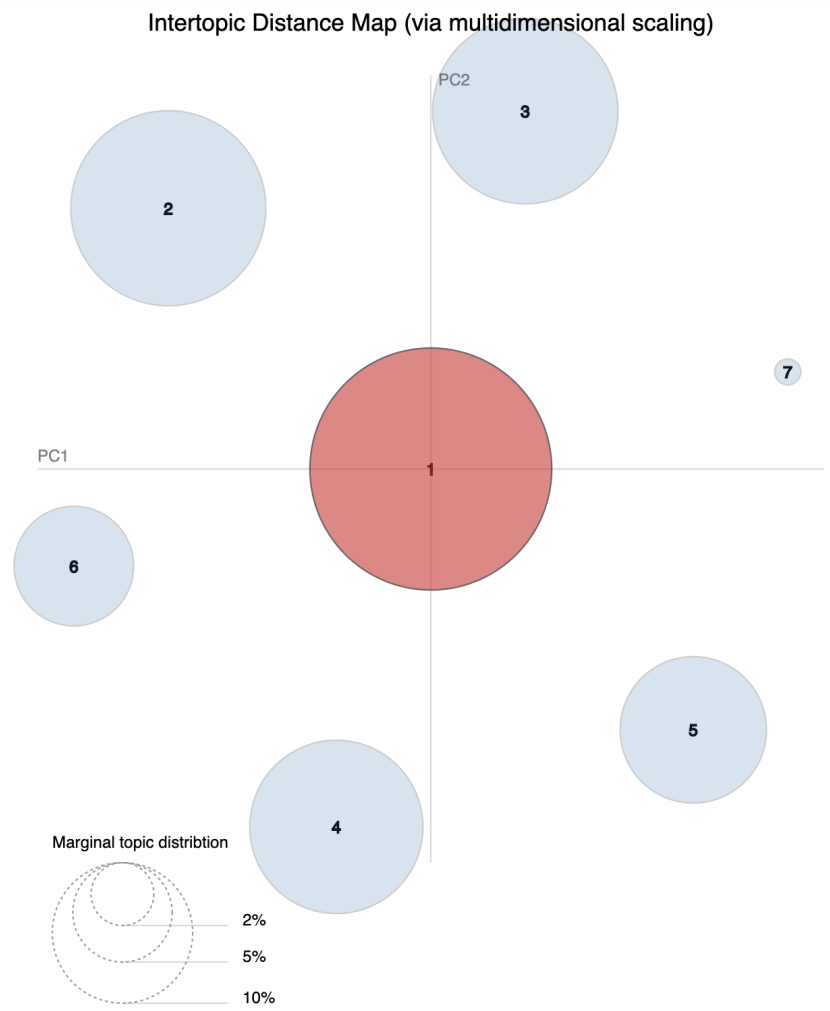
Topics from voice chats: routing

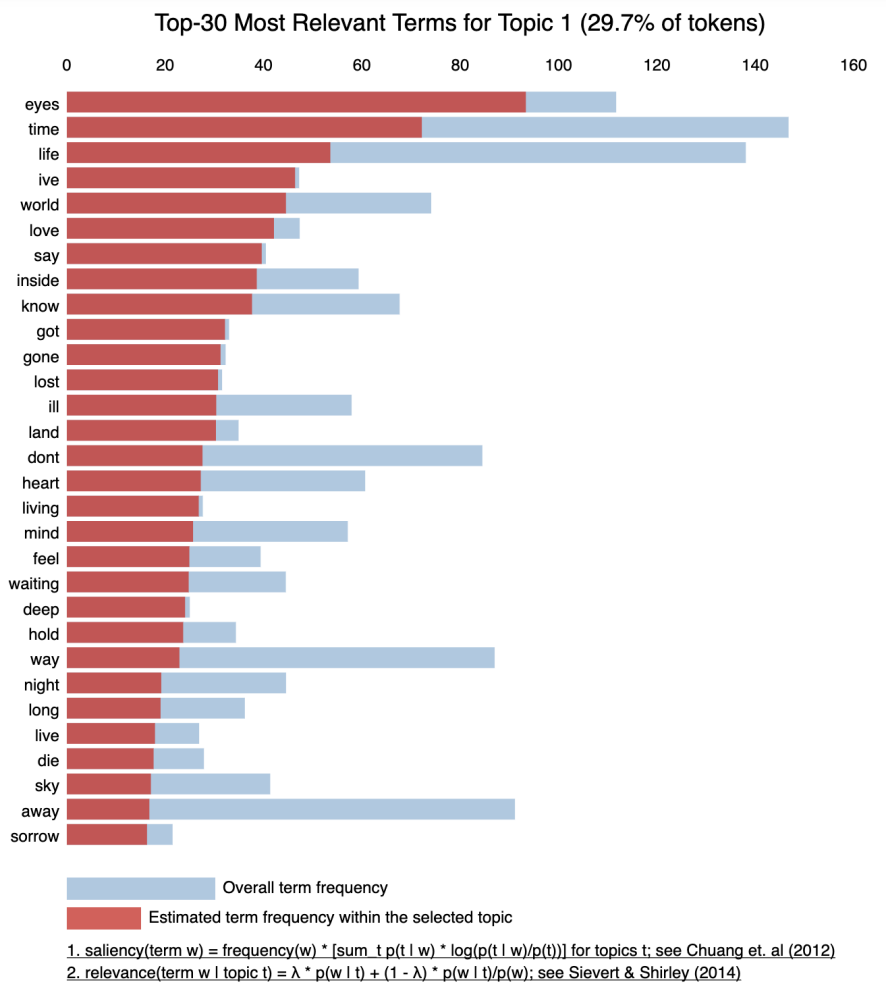
Text analysis:

- *Supervised model on top (98% precision in detecting tagged topics)*
- *For high level categories consider full text in lemmatized texts*
- *Pipeline for supervised model helpful in detailed topics*

limits	coefs
45.850572904989704	parent(bankovní produkt,platební karta)
11.270439530611739	nastavit(limit)
3.1615079323097484	dočasný(limit)
1.9202734629288676	děkovat(využití)
0.8254480537194614	levý(sloupec)
0.8135294215648389	bezpečnostní(metoda)
0.7980052161572253	ozubený(kolečko)
0.775956116424307	navýšit(limit)
0.7698551864668782	denní(limit)

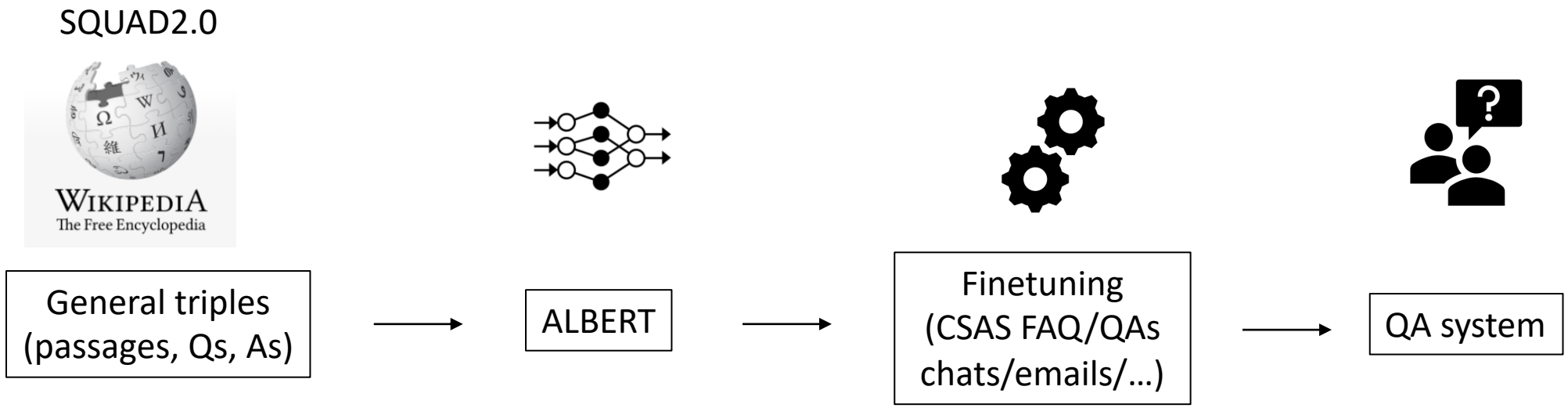
New topics detection (unsupervised clustering)



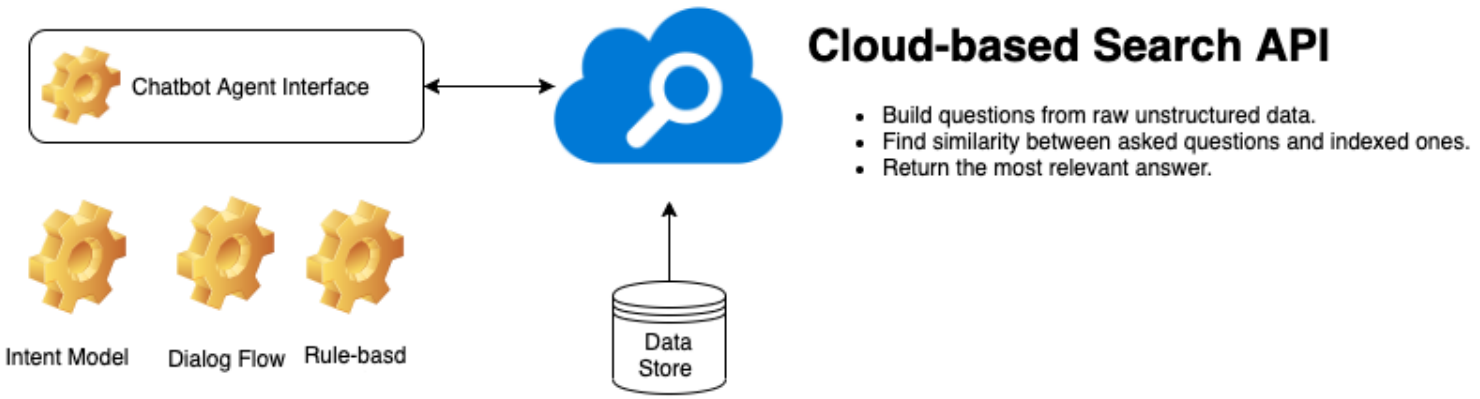


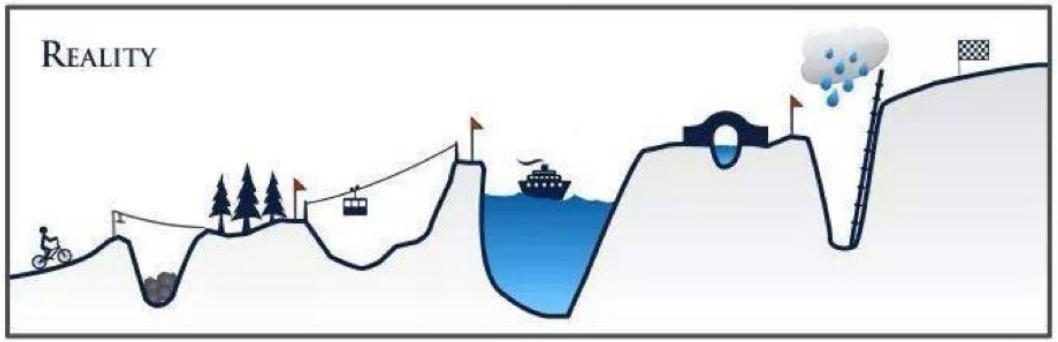
Question Extraction

- *Robust Chatbot systems need high quality Question-Answering Datasets*
- *Q&A is not a Chatbot! Powerful internal component to make chatbot smarter!*
- *Model:* ***ALBERT: A Lite BERT can be fine-tuned for Q&A task.***
- *Question-Generator is based on Sequence2sequence (deep learning).*
- *Employ search service (Azure Search, ElasticSearch ..) to index the questions and retrieve the answer on-demand.*

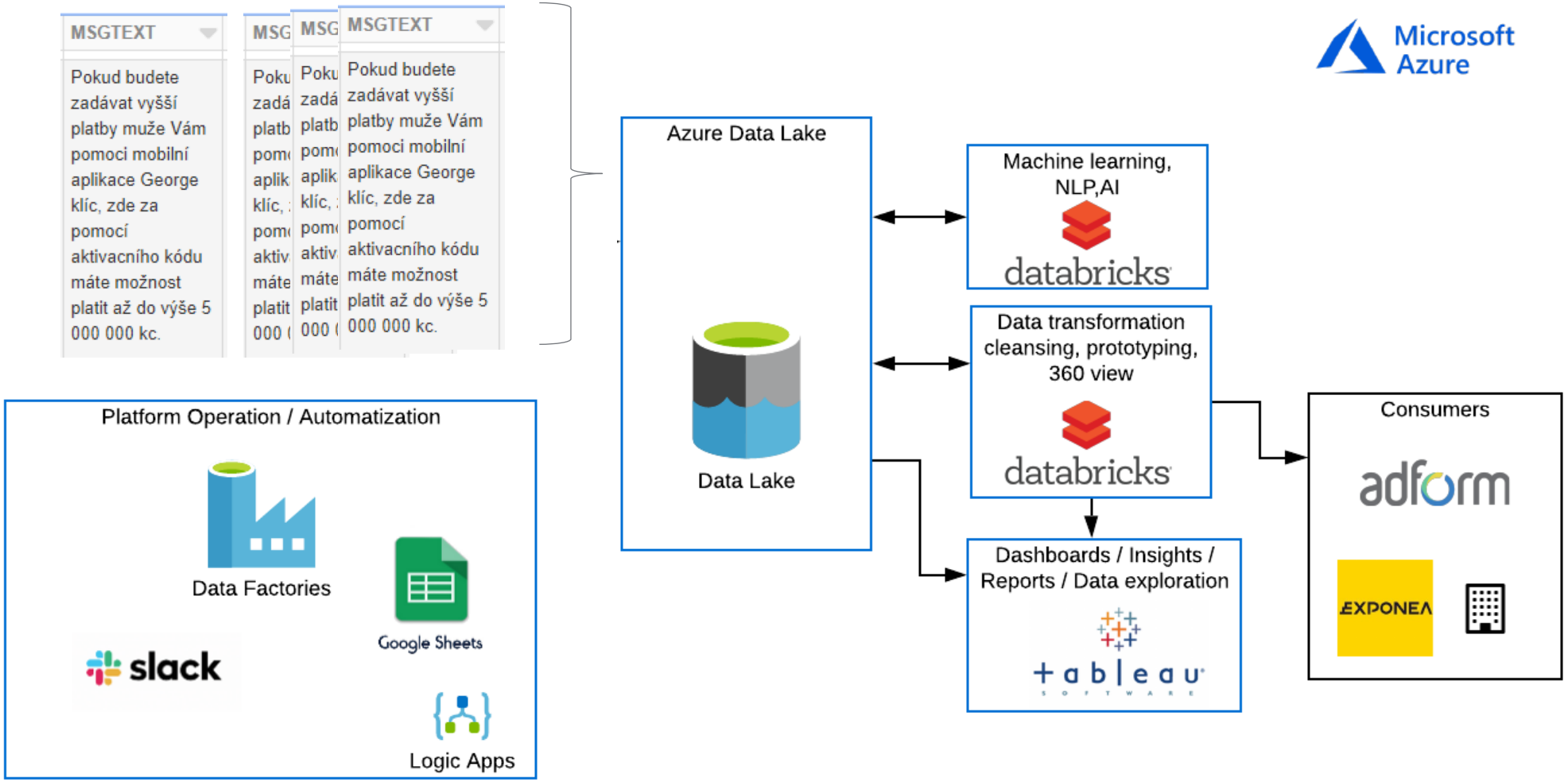


Chatbot powered by Q&A





Infrastructure



DataSentics

Contact

Jakub Stech,
Data Science architect

Washingtonova 17/1599
Prague 1, 110 00
Czech Republic

+420 775 556 122

jakub.stech@datasentics.com

www.datasentics.com