Clustering Moderate-size Collections of Short Texts

Lili Kotlerman, Ido Dagan and Oren Kurland
Bar Ilan University Technion

Overview

Motivation: industrial application of clustering user interactions

State-of-the-art:

- ✓ Make short texts longer (WordNet synonyms etc.)
- ✓ Cluster bag-of-words vectors of the enriched texts

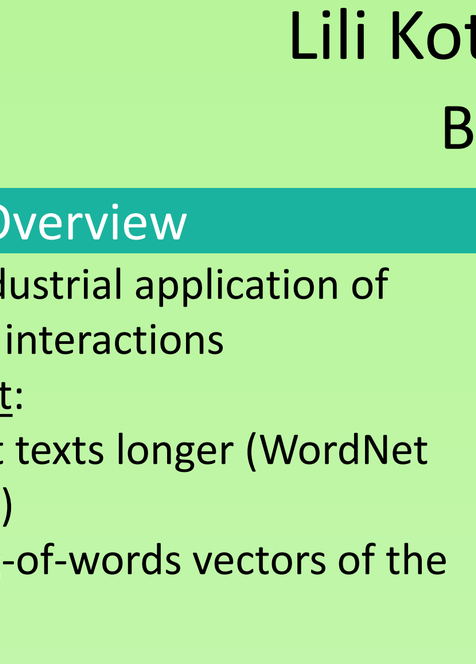
Our method:

- ✓ Perform clustering on the term level
- ✓ Project the texts over the resulting term clusters

Empirical Observations

Cluster Identifiers in Text

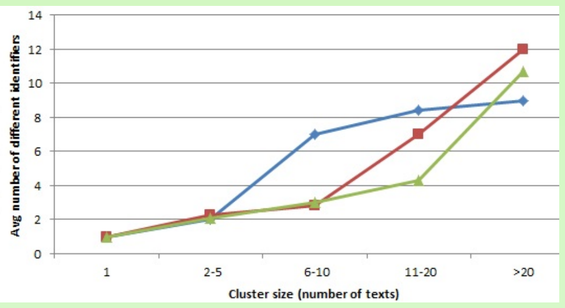
	Avg identifiers per text
Bank	1.15
Airline	1.18
Railway	1.05



Text	Cluster
it would be better if there were staff walking around selling ***snacks***	"Food" (Railway domain)
the departure date is the wrong date than what i had ***booked***	"Booking" (Airline domain)
would be ready to pay higher price for full ***dinner*** with hot ***meal***	"Food" (Railway domain)

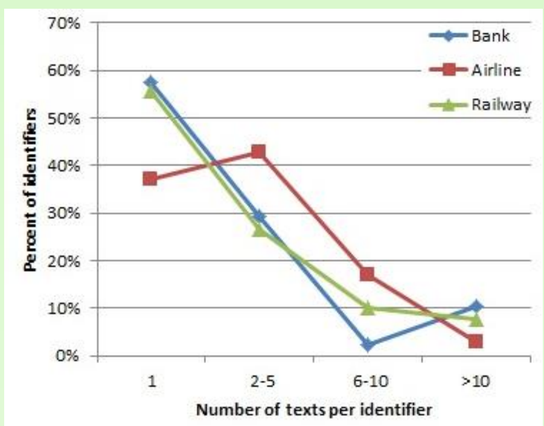
Language Variability

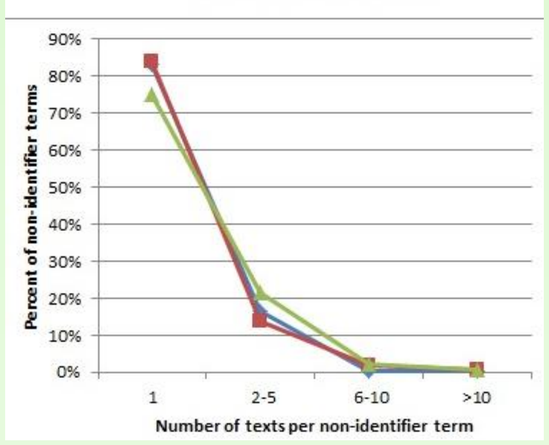
	Avg identifiers per cluster
Bank	5.76
Airline	3.58
Railway	3.24



Cluster	Identifiers
Fees and Charges (26 texts)	charge, cost, dollar, fare, fee, pay, penalize, penalty, price, spend, surcharge (11 identifiers, avg. 1.5 per text)

Distribution

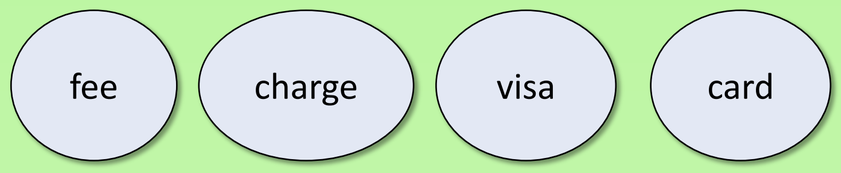




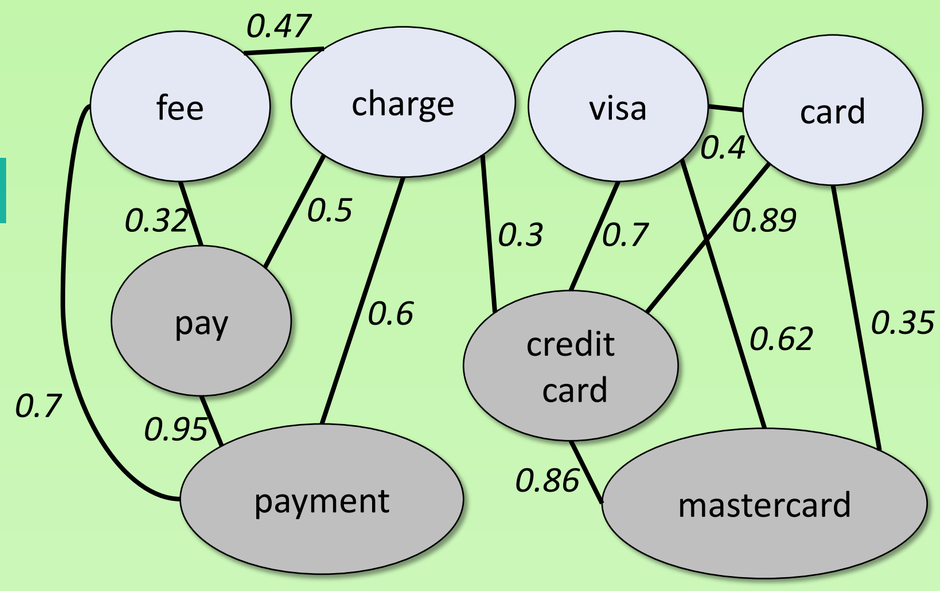
Text Clustering via Explicit Term Clusters

Step I: Generate term clusters

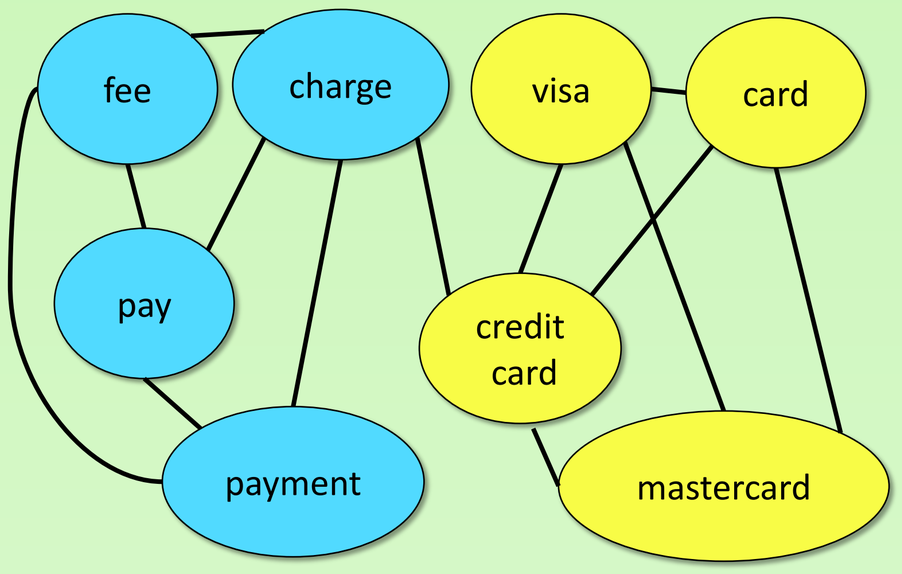
Start with the original terms



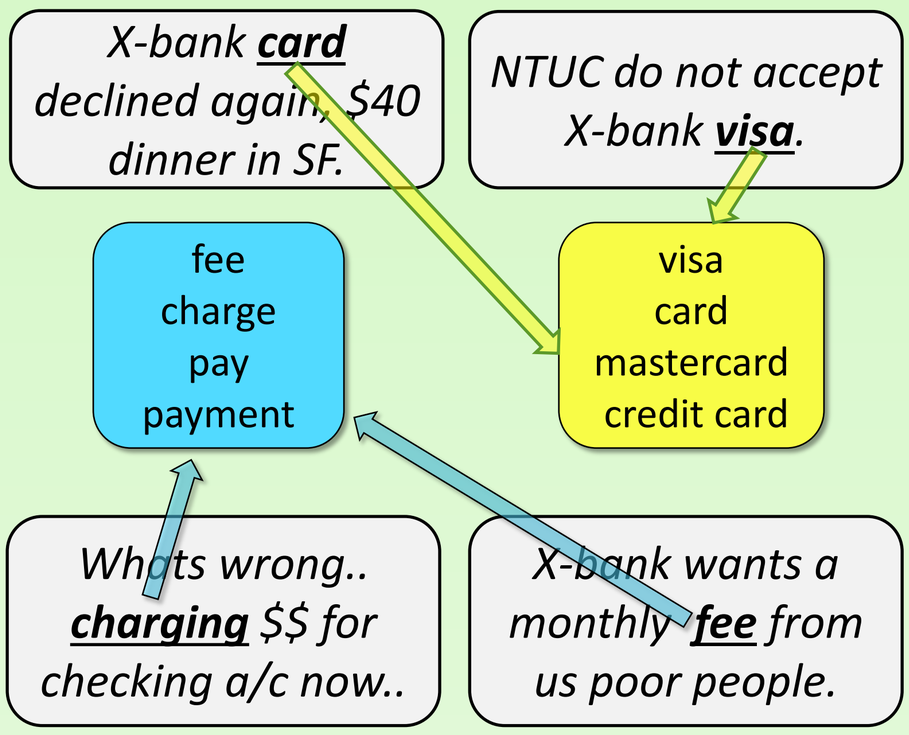
Augment them with related terms and define distances between the terms



Partition the set into term clusters



Step II: Projection over term clusters



- ✓Texts are bag-of-clusters vectors (features are term clusters like above)
- ✓Use TF-DF rather than TF-IDF weighting
- ✓Truncate the vectors to top-F features

Results (pairwise F1)

State-of-the-art methods

	Railway dataset		Bank dataset	
	TF-IDF	TF-DF	TF-IDF	TF-DF
CL-BOW	.36	.42	.21	.24
CL-BOW+	.45	**.53**	.25	**.27**
KM-BOW	.65	**.66**	.32	**.34**
KM-BOW+	.65	**.66**	.32	**.34**
Cocl-BOW	.19	.27	.10	.18
Cocl-BOW+	.31	**.37**	.11	**.20**
LDA-loc-BOW	.21	.22	.13	.26
LDA-loc-BOW+	.55	.60	.30	**.32**
LDA-ext-BOW	.15	.25	.13	.17
LDA-ext-BOW+	.51	.52	.25	.24
LDA-ext1000-BOW	.34	.28	.18	.19
LDA-ext1000-BOW+	.63	**.64**	.28	**.32**

	Airline dataset		Iphone dataset	
	TF-IDF	TF-DF	TF-IDF	TF-DF
CL-BOW	.28	**.41**	.39	.63
CL-BOW+	**.41**	.40	.68	**.69**
KM-BOW	.29	**.45**	.49	**.77**
KM-BOW+	.32	.44	.53	**.77**
Cocl-BOW	.14	.17	.15	.16
Cocl-BOW+	.18	**.22**	.18	**.31**
LDA-loc-BOW	.26	.28	.11	.18
LDA-loc-BOW+	.38	**.44**	.18	.64
LDA-ext-BOW	.16	.17	.13	.16
LDA-ext-BOW+	.33	.29	.31	.70
LDA-ext1000-BOW	.24	.26	.16	.33
LDA-ext1000-BOW+	.42	.34	**.74**	.73

Text clustering via explicit term clusters

	Railway dataset		Bank dataset	
	TF-IDF	TF-DF	TF-IDF	TF-DF
KM-BOW	.65	.66	.32	.34
KM-BOW+	.65	.66	.32	.34
KM-BOC(KM)	.51	**.71**	.38	.42
KM-BOC(KM)-top1	.56	**.71**	.26	**.49**
	TF-IDF	TF-DF	TF-IDF	TF-DF
CL-BOW	.19	.42	.21	.24
CL-BOW+	.45	.53	.25	.27
CL-BOC(KM)	.36	.45	.22	.25
CL-BOC(KM)-top1	.53	**.71**	.26	**.49**

	Airline dataset		Iphone dataset	
	TF-IDF	TF-DF	TF-IDF	TF-DF
KM-BOW	.29	.45	.49	.77
KM-BOW+	.32	.44	.53	.77
KM-BOC(KM)	.38	.39	.30	**.83**
KM-BOC(KM)-top1	.38	**.49**	.06	**.83**
	TF-IDF	TF-DF	TF-IDF	TF-DF
CL-BOW	.28	.41	.39	.63
CL-BOW+	.41	.40	.68	.69
CL-BOC(KM)	.35	.39	.43	.53
CL-BOC(KM)-top1	.38	**.49**	.06	**.83**

Conclusions

- ✓ Four new datasets (will be published)
- ✓ Analysis of the datasets with clustering in mind
- ✓New method, improved performance
- ✓Clustering tool will be made publicly available