Event Detection from Text Data

Tomáš Kala

June 20, 2017

Event detection

- What is it about?
- Original method by He et al. (2007)
- Our contribution (through Word2Vec)

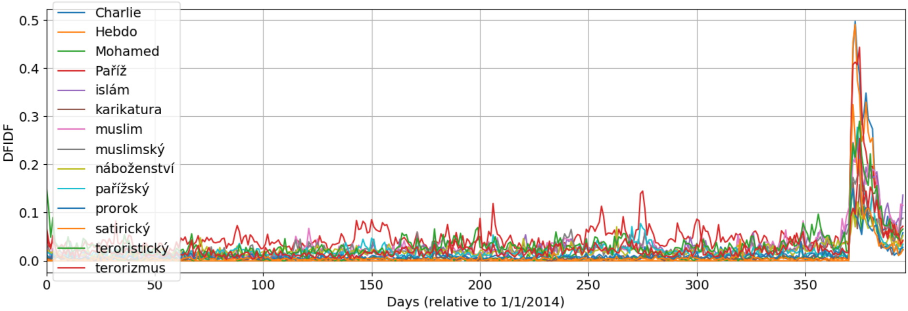


Figure: 6/1 - 17/1, 2015: V redakci satirického listu Charlie Hebdo v Paříži se střílelo. Francouzský satirický časopis Charlie Hebdo, na který minulý týden zaútočili islamisté, znovu vydá karikatury proroka Mohameda.

Word2Vec

- Neural network language model by Mikolov et al. (2013)
- Learns vector representation that preserves word properties

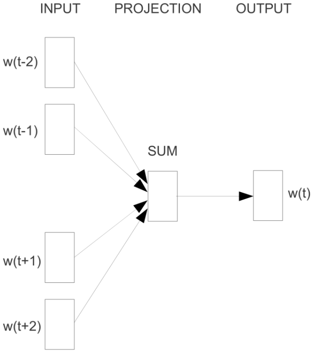


Figure: Word2Vec schema

terorista	**olympiáda**
islamista	olympijský
džihádista	paralympiáda
extremista	univerziáda
teroristický	Soča
Coulibaly	medailista
allah	Soči
ozbrojenec	víceboj
džihád	mistrovství
islámský	šampionát

Table: Most similar words

Word representation

Each word abstracted into 2 vectors

1. Semantical representation – vector space embedding

$$\mathbf{v}_w \in \mathbb{R}^{100} \text{ (learned through Word2Vec).} \tag{1}$$

2. Trajectory – Document Frequency-Inverse Document Frequency

$$\mathbf{y}_w \in \mathbb{R}^T, \ \mathbf{y}_w(t) = \underbrace{\frac{\mathrm{DF}_w(t)}{\mathrm{N}(t)}}_{\mathrm{DF}} \cdot \underbrace{\log \frac{N}{\mathrm{DF}_w}}_{\mathrm{IDF}}, \ t = 1, \ldots, T \tag{2}$$

Word trajectories

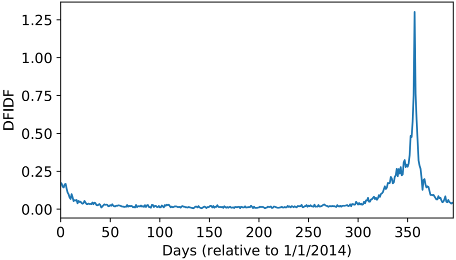


Figure: An important word (Christmas)

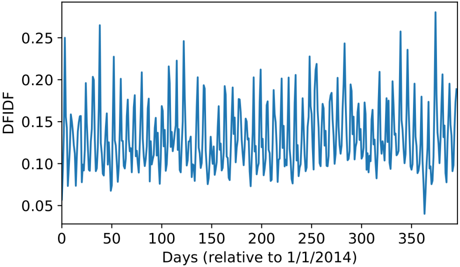


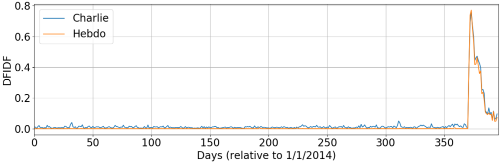
Figure: A stopword (Friday)

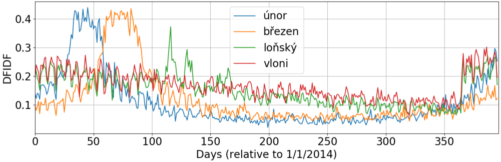
Signal power decides between the two categories.

Event detection

Original method and its modification

1. Original greedy optimization:
   - KL-divergence of the trajectories
   - Simple document overlap
   - 217 events, 2.08 keywords/event
   - Too strict
2. Word2Vec-based modification:
   - Cosine similarity of word vectors
   - 46 events, 10.28 keywords/event
   - Too noisy

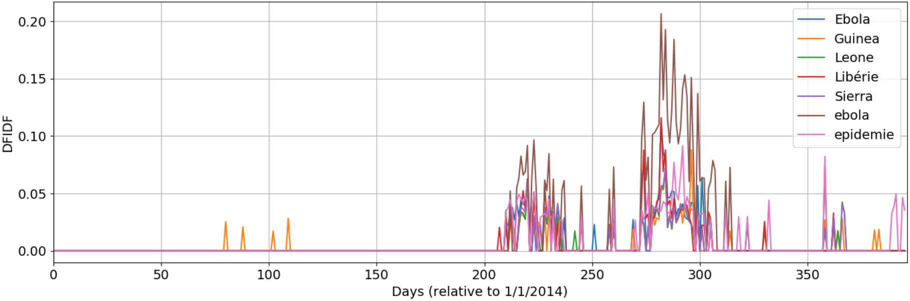




Event detection

Cluster-based algorithm

- Application of DBSCAN (Ester et al., 1996)
- Custom distance function
- Trajectory filtering
- 77 events, 9.88 keywords/event



Document retrieval

- Event trajectories
- Active periods
- Keywords as a query

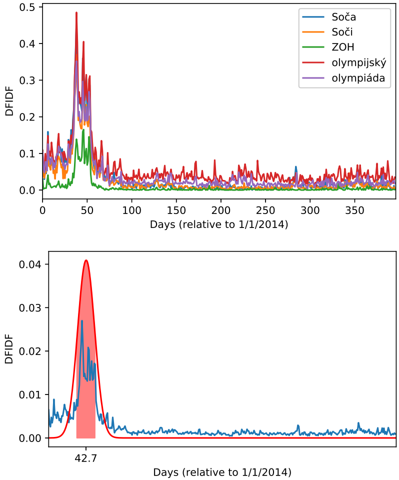


Figure: Gaussian fit, active period = $[\mu - \sigma, \mu + \sigma]$

Event annotation

Document headlines not informative enough

Charlie Hebdo opět otiskne karikatury proroka Mohameda

Multi-document summarization (Lin and Bilmes, 2010, 2011)

$$\begin{aligned} \max_{S \subseteq U} \quad & \mathcal{F}(S) = \mathcal{L}(S) + \lambda \mathcal{R}(S) \\ \text{s. t.} \quad & \sum_{i \in S} c_i \leq \mathcal{B} \end{aligned} \tag{3}$$

We ran into some issues...

. . . Pak ale začalo zabíjení v centru Paříže. . . . Sloni v zoo Dvůr Králové si pochutnali na vanočních stromcích. . . .

Results

Method	P	R	F_1	Redundancy	Noisiness	Purity
Original	16.35%	**28.57%**	20.80%	77.99%	50.94%	30.53%
Modified	8.70%	10.20%	9.39%	65.22%	19.57%	44.42%
Clusters	**25.97%**	**28.57%**	**27.21%**	**42.86%**	**19.48%**	**61.08%**

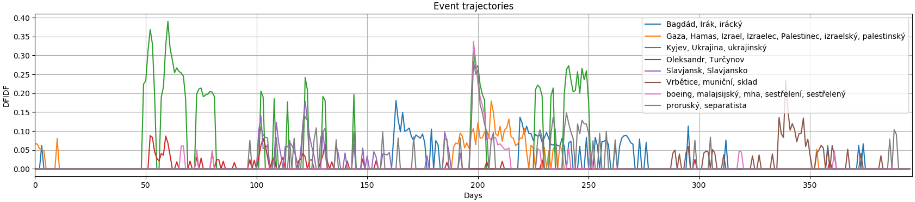
Table: Precision, Recall, Redundancy, Noisiness and Purity comparison

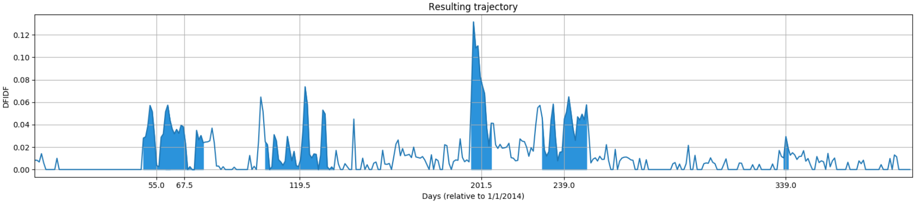
Unit	Original	Modified	Clusters
Word2Vec	N/A	3h 50min	
Word analysis	⟵	37min	⟶
Event detection	2min 12s	38s	4min 50s
Document retrieval	7min 30s	6h	7h 40min
Event annotation	3h 22min	3min 38s	7min 30s
Total	4h 9min	10h 31min	12h 20min

Table: Computation time comparison

Conclusion, use case

- Event detection with low redundancy
- Subsequent document retrieval
- Human-readable annotations





Bibliography

M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

Q. He, K. Chang, and E.-P. Lim. Analyzing feature trajectories for event detection. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 207–214, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-597-7. doi: 10.1145/1277741.1277779. URL `http://doi.acm.org/10.1145/1277741.1277779`.

M. J. Kusner, Y. Sun, N. I. Kolkin, K. Q. Weinberger, et al. From word embeddings to document distances. In *ICML*, volume 15, pages 957–966, 2015.

H. Lin and J. Bilmes. Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 912–920. Association for Computational Linguistics, 2010.

H. Lin and J. Bilmes. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 510–520. Association for Computational Linguistics, 2011.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. URL `http://arxiv.org/abs/1301.3781`.

Word representation

Each word abstracted into 2 vectors

1. Semantical representation – vector space embedding

$$\mathbf{v}_w \in \mathbb{R}^{100} \text{ (learned through Word2Vec).} \tag{4}$$

2. Trajectory – Document Frequency-Inverse Document Frequency

$$\mathbf{y}_w \in \mathbb{R}^T,\ \mathbf{y}_w(t) = \underbrace{\frac{\mathrm{DF}_w(t)}{\mathrm{N}(t)}}_{\mathrm{DF}} \cdot \underbrace{\log \frac{N}{\mathrm{DF}_w}}_{\mathrm{IDF}},\ t = 1, \ldots, T \tag{5}$$

with

- T . . . document stream length (in days),
- N . . . number of documents,
- $\mathrm{N}(t)$ . . . # of documents published on day t,
- DF_w . . . # of documents containing the word w,
- $\mathrm{DF}_w(t)$ . . . # of documents containing the word w published on day t.

Multi-document summarization

$$\max_{S \subseteq U} \quad \mathcal{F}(S) = \mathcal{L}(S) + \lambda \mathcal{R}(S)$$
$$\text{s. t.} \quad \sum_{i \in S} c_i \leq \mathcal{B}, \text{ with} \tag{6}$$

- U . . . set of all sentences,
- $\mathcal{L}$ . . . relevance measure composed of sentence pairwise similarities,
- $\mathcal{R}$ . . . diversity measure controlled by λ,
- $\mathcal{B}$ . . . maximum summary length,
- c_i . . . length of sentence i.

Word Mover's Distance

- Document similarity measure by Kusner et al. (2015)
- Transportation problem between word vectors of 2 documents

$$
\begin{aligned}
\min_{\mathbf{T}\geq 0} \quad & \sum_{i,j=1}^{n} \mathbf{T}_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2 \\
\text{s. t.} \quad & \sum_{j=1}^{n} \mathbf{T}_{ij} = d_i \quad \forall i \in \{1, \ldots, n\} \\
& \sum_{i=1}^{n} \mathbf{T}_{ij} = d'_j \quad \forall j \in \{1, \ldots, n\}
\end{aligned}
\tag{7}
$$

- n . . . vocabulary size
- $\mathbf{x}_i$ . . . vector embedding of the word i
- d_i (d'_i) . . . normalized frequency of i in document 1 (2)

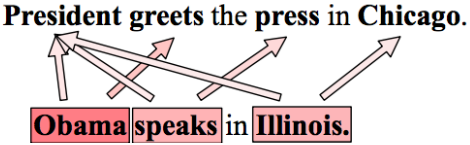

President greets the press in Chicago.
Obama speaks in Illinois.

Figure: WMD illustration