Text Summarization
29 Ağustos - 3 Eylül 18

Text summarization is the task of shortening the content of a piece of text while retaining the story and coherence of it. There are two main approaches; extractive and abstractive summarization, namely. In extractive approaches, only the heaviest, the most meaningful, the most informative sentences of the text are selected as a subset. In the abstractive approach, the theme of the text, after captured, is conveyed with different sentences and a shorter, new piece of text is output. While abstractive approach is more difficult and prone to producing irrelevant words, extractive summarization might result in a bunch of disconnected sentences if the sentence selection criteria are not well-set. Language independency and easy implementation as well as giving acceptable output make extractive summarization more widely used and studied.

Two surveys on summarization, one introducing the field [1] and the other focusing more on recent studies [2] are available to better know the methods. A recent survey on summarization techniques for Turkish language can be found in [3].

Method

In this work, we apply extractive approach to summarize texts in Turkish. The method is called textrank [4], which is based on pagerank, building a graph out of the sentences in the text and determining the most important ones with their in-degree and out-degree ratios. Given a text, it is first split into sentences and these sentences are represented as vectors which contain tfidf weights of all the words in the text. The matrix made up of sentence vectors is multiplied with its transpose to get a similarity matrix of the sentences; thus, those with more common words having higher weights will have higher similarity values in the similarity matrix. This matrix can easily be considered an undirected graph with nodes as sentences and the edge weights as similarity values. Applying pagerank to find the most important nodes, the sentences are sorted in the end and the top N ones are output as the summary elements.

Although there are many studies on summarization for Turkish with useful datasets, we couldn’t get any of these datasets; the authors do not respond to the emails. And we cannot present a performance evaluation as we yet don’t have a gold set. Performance of textrank on English data can be found on the source paper [4].

Results

We also run summarization experiments with LSA (Latent Semantic Analysis) [5] but didn’t see remarkable / more acceptable summaries.

Here we give examples for Turkish and English.

English

Original text	Summary
It seems self-evident: an overworked person is tired; hence more likely to have an accident at work. But proving this is surprisingly difficult. It might be that riskier jobs also have more demanding hours, or simply that people who work more hours spend more time at risk, even if they don’t do overtime. But a study that analysed 13 years of job records in the US found that “working in jobs with overtime schedules was associated with a 61% higher injury hazard rate compared to jobs without overtime”. This specific study stops short of saying that fatigue is the primary cause of this increased risk, but there is ample evidence to suggest this might be the case. It makes accidents more likely, boosts stress levels, and even causes physical pain. But the real problem is that many people just can’t afford not to do it. According to latest International Labour Organization statistics, more than 400 million employed people worldwide work 49 or more hours per week, a sizeable proportion of the near 1.8 billion total employed people worldwide. But wearing exhaustion like a badge of honour sets a dangerous precedent. Hustling over long hours and weekends has become a staple of start-up culture in Silicon Valley - hence, it has also filtered out to many parts of the world. The problem is that this 'long hours' culture likely defeats the purpose of getting more things done, or at least puts a very hefty price on doing them.	It might be that riskier jobs also have more demanding hours, or simply that people who work more hours spend more time at risk, even if they don’t do overtime. This specific study stops short of saying that fatigue is the primary cause of this increased risk, but there is ample evidence to suggest this might be the case. But the real problem is that many people just can’t afford not to do it. The problem is that this 'long hours' culture likely defeats the purpose of getting more things done, or at least puts a very hefty price on doing them.

Turkish

Original text	Summary
Tek tek her kişide, başkalarının vereceği zarardan korunan kişisel nitelik ve eylemlere hak adını veren, hakkı “bireyin başka herkes karşısında olumlayıp, öne sürdüğü ve koruduğu bir şey olma veya bir şey yapma ya da bir şeye sahip olma özgürlüğü” olarak tanımlayan Smith’te “haklar”, “zarar” ve “ahlaki kişilik” arasında yakın bir ilişki bulunur. Çünkü ona göre, muhayyile sosyal deneyime bağlı olup, bir sosyal düzeyden diğerine farklılık gösterdiği için hakların ve ahlaki kişiliğin kendisi de bir toplum biçiminden bir başkasına farklılık gösterir. Onun sadece ahlaklılığa değil adalet ve hukuka yönelik tarihsel yaklaşımı da işte burada karşımıza çıkar. Gerçekten de insan doğasını anlamak için kullanılacak bir araç olarak “doğa durumu” düşüncesini reddeden ve insana insanlığını kazandıran şeyin kendisine tutulan sosyal ayna olduğu için insan türünün moral hayatının kaçınılmaz olarak toplumsal olduğunu öne süren Smith’e göre, sadece ahlaki özellik ve erdemlerimizle değil fakat haklarla ilgili değerlendirmelerin de sosyal bir çerçeveye oturtulması gerekir. Dört Aşamalı Sosyal Gelişme Teorisi Smith, işte bu çerçeve içinde toplumsal gelişmenin dört ayrı evresini kişilik kavramındaki genişleme ve hakların kapsamında kaydedilen ilerleme bakımından birbirinden ayırır. Onun aynı zamanda “dört evreli bir tarihsel ilerleme” anlayışından oluşan bu tarih felsefesi görüşü, Aydınlanmanın veya Aydınlanma filozoflarının tarihe bilimsel bakışlarının bir başka ifadesi olmasının dışında, ilerlemenin temelinde entelektüel gelişme ve bilimsel bilgi birikiminden ziyade iktisadi faktörlerin olduğunu öne sürmek açısından da büyük önem taşır. Benzerlerini sadece Aydınlanma	Gerçekten de insan doğasını anlamak için kullanılacak bir araç olarak “doğa durumu” düşüncesini reddeden ve insana insanlığını kazandıran şeyin kendisine tutulan sosyal ayna olduğu için insan türünün moral hayatının kaçınılmaz olarak toplumsal olduğunu öne süren Smith’e göre, sadece ahlaki özellik ve erdemlerimizle değil fakat haklarla ilgili değerlendirmelerin de sosyal bir çerçeveye oturtulması gerekir. Onun aynı zamanda “dört evreli bir tarihsel ilerleme” anlayışından oluşan bu tarih felsefesi görüşü, Aydınlanmanın veya Aydınlanma filozoflarının tarihe bilimsel bakışlarının bir başka ifadesi olmasının dışında, ilerlemenin temelinde entelektüel gelişme ve bilimsel bilgi birikiminden ziyade iktisadi faktörlerin olduğunu öne sürmek açısından da büyük önem taşır. Benzerlerini sadece Aydınlanma filozoflarında değil, fakat Hegel, Marx ve Comte gibi 19. yüzyıl filozoflarından da görebileceğimiz bu ilerleme ya da sosyal gelişme teorisi tarihte, haklar, özgürlük, kişilik bakımından gerçek bir ilerleme olduğu varsayımına dayanır.

filozoflarında değil, fakat Hegel, Marx ve Comte gibi 19. yüzyıl filozoflarından da görebileceğimiz bu ilerleme ya da sosyal gelişme teorisi tarihte, haklar, özgürlük, kişilik bakımından gerçek bir ilerleme olduğu varsayımına dayanır.	

We can work on many improvements in time; for example, better preprocessing like the removal of stopwords; using word embeddings instead of tfidf weighted vectors to include more of real world meanings of words; applying improved trials of textrank.

References

[1] Nenkova, Ani, and Kathleen McKeown. "A survey of text summarization techniques." *Mining text data*. Springer, Boston, MA, 2012. 43-76.
[2] Yao, Jin-ge, Xiaojun Wan, and Jianguo Xiao. "Recent advances in document summarization." Knowledge and Information Systems 53.2 (2017): 297-336.
[3] Birant, Çağdaş Can, Özgün Koşaner, and Özlem Aktaş. "A Survey to Text Summarization Methods for Turkish." *International Journal of Computer Applications* 144.6 (2016).
[4] Mihalcea, Rada. "Graph-based ranking algorithms for sentence extraction, applied to text summarization." *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. Association for Computational Linguistics, 2004.
[5] Landauer, Thomas K., and Susan T. Dumais. "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological review* 104.2 (1997): 211.