

## Methods

We applied three methods:

- 1) Baseline: Simple chunk extraction
- 2) Tfidf weighting: Weight unigrams and chunks in the text; output the first N heaviest as the keyphrase
  - This method actually requires a set of texts to get the score of phrases weighted throughout other texts. For single-text keyphrase extraction, we add a couple of auxiliary short texts.
- 3) Generic keyphrase extraction - adapted from [1]: This method is based on mutual information of words taking their word embeddings into that weighting as well and can work for single texts. POS tagging is not applied for the sake of quick processing but would be better if applied.
  - For now, it works only on word pairs.

These three methods are ready for english and turkish; can easily be applied for french or other languages.

## Results

On Turkish texts

**Text** = “Üzerinde duman tüten yıkıntı dağları, parçalanmış ceset yığınları, baktığınız her yerde buhar ve duman tüten bir ateş denizi, çamur ve küller; çırpınan bir kırlangıç gibi volkanın kayalık yamacına konmuş kıpır kıpır küçük şehirden tüm kalanlar işte bunlar. Öfkeli dev, bu insan cüretine, iki bacaklı cücelerin kör kendini beğenmişliğine karşı bir süredir gürleyip köpürüyordu. Gazabında bile iyi yürekli, vefalı olan bu dev, ayaklarına çıkmış sürünen bu fütursuz yaratıkları uyarıyordu. Dumanlar çıkarıyor, ateşten bulutlar kusuyordu, bağrında fokurtular, kaynamalar ve tüfek mermileri ve top gümbürtüsü gibi patlamalar oluyordu. Fakat insanın kaderine hükmeden yeryüzünün efendileri, kendi bilgeliklerine sarsılmaz bir inanç duyuyorlardı.”

## Keyphrases

Baseline - chunks	Tfidf weighted chunks and unigrams (sorted by phraseness score)	Keyphrases (only pairs) extracted with generic method of Wang et al. 2015 (sorted by phraseness score)
'tüten bir ateş', 'bir ateş denizi', 'çırpınan bir kırlangıç', 'volkanın kayalık yamacına', 'sarsılmaz bir inanç', 'üzerinde duman', 'tüten yıkıntı', 'yıkıntı dağları', 'parçalanmış ceset', 'ceset yığınları', 'kıpır kıpır', 'küçük şehirden', 'insan cüretine', 'bacaklı cücelerin', 'tüfek mermileri', 'top gümbürtüsü', 'insanın kaderine', 'yeryüzünün efendileri', 'kendi bilgeliklerine'	'duman', 'ateş', 'dev', 'insan', 'kıpır', 'tüt', 'ayak', 'bacak', 'bacaklı cücelerin', 'bak', 'bağır', 'beğen', 'bilge', 'bir ateş denizi', 'buhar', 'bulut', 'ceset', 'ceset yığınları', 'cüce', 'cüret', 'dağ', 'deniz', 'efendi', 'fütur', 'gazap', 'gümbürtü', 'inanç', 'insan cüretine', 'insanın kaderine',	('Üzerinde', 'duman'), ('ateş', 'denizi'), ('konmuş', 'kıpır'), ('ve', 'tüfek'), ('gibi', 'patlamalar'), ('bilgeliklerine', 'sarsılmaz'), ('kaderine', 'hükmeden'), ('patlamalar', 'oluyordu'), ('gibi', 'volkanın'), ('yeryüzünün', 'efendileri'), ('bulutlar', 'kusuyordu'), ('sarsılmaz', 'bir'), ('ceset', 'yığınları'), ('ayaklarına', 'çıkmış'), ('bir', 'süredir'), ('cücelerin', 'kör'), ('Fakat', 'insanın'), ('şehirden', 'tüm'), ('parçalanmış', 'ceset'), ('yıkıntı', 'dağları'), ('bacaklı', 'cücelerin'), ('küçük', 'şehirden'), ('Dumanlar', 'çıkarıyor'), ('iyi', 'yürekli'), ('beğenmişliğine', 'karşı'), ('bile', 'iyi'), ('fütursuz', 'yaratıkları'), ('yaratıkları', 'uyarıyordu'), ('her', 'yerde'), ('kayalık', 'yamacına')

- Chunks seem the best set of phrases and the generic methods' output appear promising. Considering the quality and domain-specificity, generic method might give better results upon such needs.

## On English texts

**Text** = "Mountains of smoking ruins, heaps of mangled corpses, a steaming, smoking sea of fire wherever you turn, mud and ashes – that is all that remains of the flourishing little city which perched on the rocky slope of the volcano like a fluttering swallow. For some time the angry giant had been heard to rumble and rage against this human presumption, the blind self-conceit of the two-legged dwarfs. Great-hearted even in his wrath, a true giant, he warned the reckless creatures that crawled at his feet. He smoked, spewed out fiery clouds, in his bosom there was seething and boiling and explosions like rifle volleys and cannon thunder. But the lords of the earth, those who ordain human destiny, remained with faith unshaken – in their own wisdom."

## Keyphrases

Baseline - chunks	Tfidf weighted chunks and unigrams (sorted by phraseness score)	Keyphrases (only pairs) extracted with generic method of Wang et al. 2015 (sorted by phraseness score)
'mountains of smoking ruins', 'heaps of mangled corpses', 'steaming', 'sea of fire', 'ashes', 'flourishing little city', 'rocky slope', 'volcano', 'fluttering swallow', 'time', 'angry giant',	'bosom', 'earth', 'feet', 'giant', 'human', 'time', 'volcano', 'wrath', 'angri', 'angry giant', 'ash', 'ashes', 'blind',	('rifle', 'volleys'), ('Mountains', 'of'), ('spewed', 'out'), ('who', 'ordain'), ('unshaken', '—'), ('cannon', 'thunder'), ('smoking', 'ruins'), ('human', 'destiny'), ('mangled', 'corpses'), ('human', 'presumption'), ('and', 'ashes'), ('fiery', 'clouds'), ('mud', 'and'),

'human presumption', 'blind self-conceit', 'two-legged dwarfs', 'wrath', 'true giant', 'reckless creatures', 'feet', 'fiery clouds', 'bosom', 'explosions like rifle volleys', 'cannon thunder', 'lords', 'earth', 'human destiny', 'faith unshaken –', 'own wisdom'	'blind self-conceit', 'cannon', 'cannon thunder', 'citi', 'cloud', 'corps', 'creatur', 'destini', 'dwarf', 'explos', 'explosions like rifle volleys', 'Faith', 'faith unshaken –', 'fieri', 'fiery clouds', 'flourish', 'flourishing little city', 'flutter', 'fluttering swallow', 'great-heart',	('fire', 'wherever'), ('and', 'boiling'), ('bosom', 'there'), ('seething', 'and'), ('to', 'rumble'), ('self-conceit', 'of'), ('lords', 'of'), ('perched', 'on'), ('heard', 'to'), ('that', 'crawled'), ('a', 'steaming'), ('a', 'true'), ('own', 'wisdom'), ('this', 'human'), ('remained', 'with'), ('with', 'faith'), ('the', 'lords'), ('the', 'reckless'), ('Great-hearted', 'even'), ('the', 'angry'), ('the', 'flourishing'),
--	--	---

- We can give the similar explanations we did for Turkish outputs. Preprocessing might improve the results of tfidf-based and generic methods.

[1] Wang, Rui, Wei Liu, and Chris McDonald. "Corpus-independent generic keyphrase extraction using word embedding vectors." Software Engineering Research Conference. Vol. 39. 2014.