COMPARING RESEARCH INITIATIVES USING NLP

Author: Dongho Chun - Mentor: Marcel Parciak - (Co-) supervisor: prof. Dr. Liesbet Peeters, prof. Dr. Stijn Vansummeren

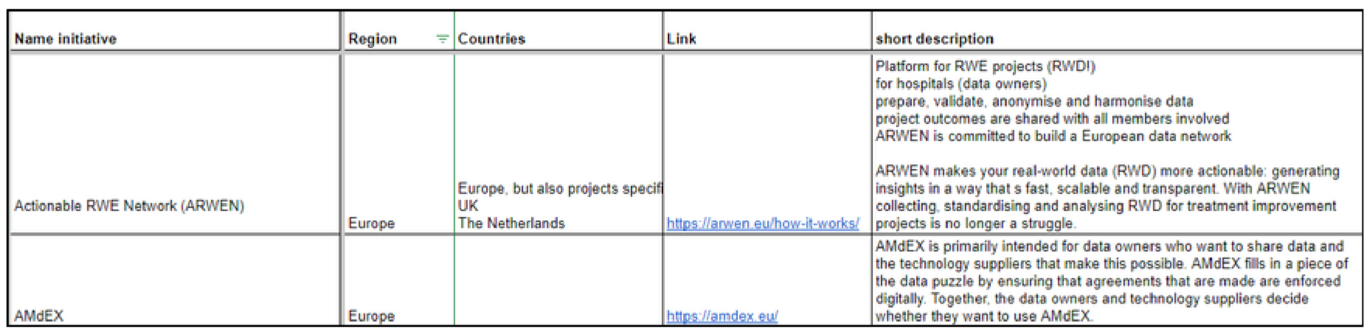
INTRODUCTION

- **Initiatives** working with health data
- Some are **closely related** to each other
- **Poor documentation**

RESEARCH QUESTIONS

- How can the initiatives be **compared** with each other?
- How can the initiatives that are closely related to each other be **identified** based on their **description**?

DATASET



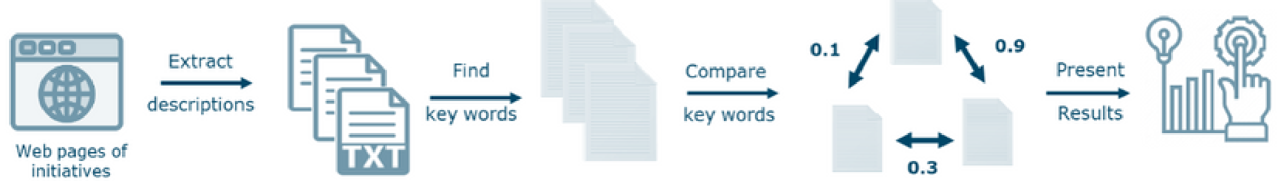
Name initiative	Region	Countries	Link	short description
Actionable RWE Network (ARWEN)	Europe	Europe, but also projects specifi UK The Netherlands	https://arwen.eu/how-it-works/	Platform for RWE projects (RWD!) for hospitals (data owners) prepare, validate, anonymise and harmonise data project outcomes are shared with all members involved ARWEN is committed to build a European data network ARWEN makes your real-world data (RWD) more actionable: generating insights in a way that s fast, scalable and transparent. With ARWEN collecting, standardising and analysing RWD for treatment improvement projects is no longer a struggle.
AMdEX	Europe		https://amdex.eu/	AMdEX is primarily intended for data owners who want to share data and the technology suppliers that make this possible. AMdEX fills in a piece of the data puzzle by ensuring that agreements that are made are enforced digitally. Together, the data owners and technology suppliers decide whether they want to use AMdEX.

- Strategic oversight of **68 initiatives**
- Short description + link

CONCLUSION

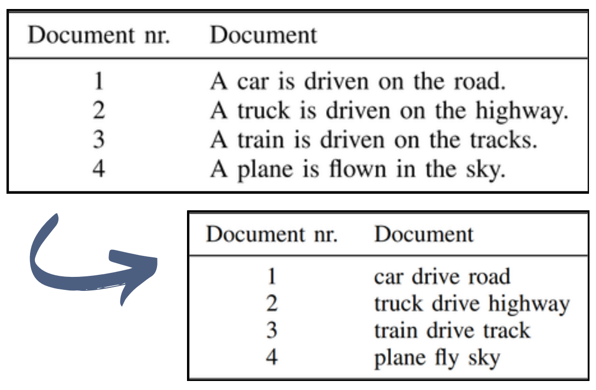
- A **tool** for **partial automation**
- **Subjectivity** in “k”
- **Wide variation** in the pipeline
- Overall **satisfactory results**

PIPELINE



Data Extraction

- Webscraper
- Lemmatization
- Stopwords removal
- Punctuation and numbers removal

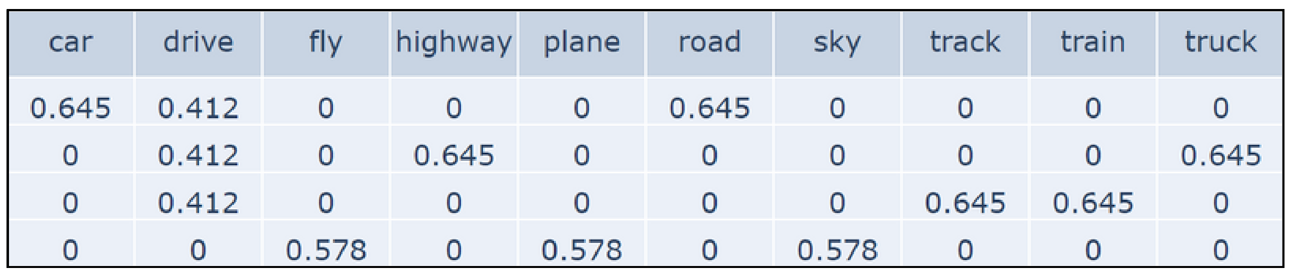


Document nr.	Document
1	A car is driven on the road.
2	A truck is driven on the highway.
3	A train is driven on the tracks.
4	A plane is flown in the sky.

Document nr.	Document
1	car drive road
2	truck drive highway
3	train drive track
4	plane fly sky

Finding Keywords

TF-IDF



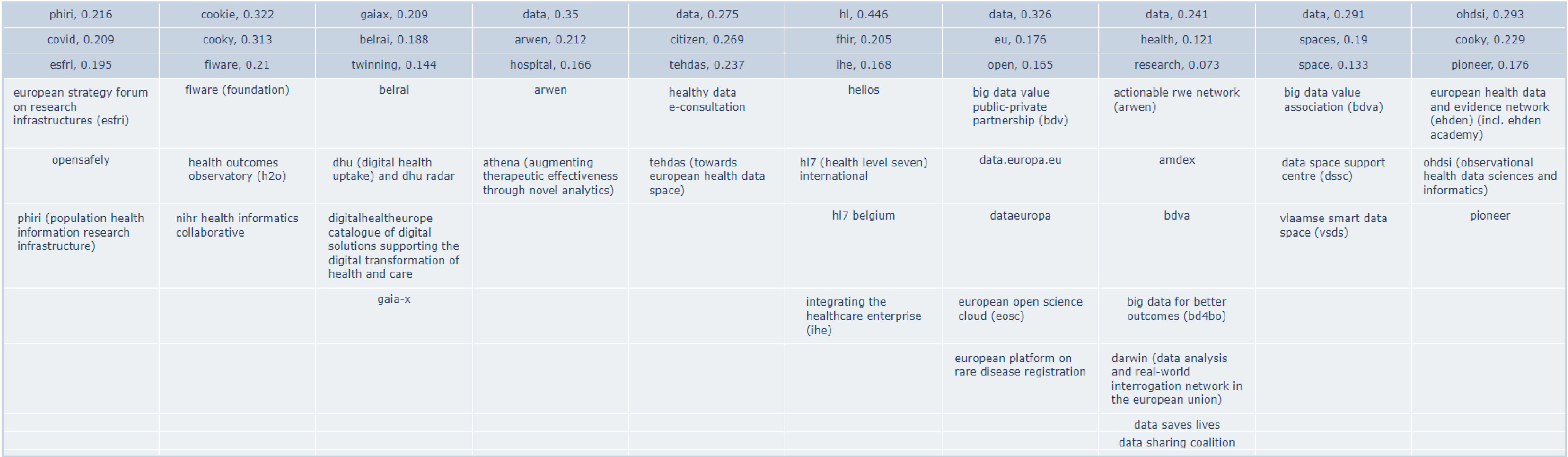
car	drive	fly	highway	plane	road	sky	track	train	truck
0.645	0.412	0	0	0	0.645	0	0	0	0
0	0.412	0	0.645	0	0	0	0	0	0.645
0	0.412	0	0	0	0	0	0.645	0.645	0
0	0	0.578	0	0.578	0	0.578	0	0	0

Comparing Keywords

K-means clustering

Visualizing Results

- 1 column = 1 cluster
- Similar initiatives grouped together
- Column header = top 3 words + TF-IDF



phiri, 0.216	cookie, 0.322	gaiax, 0.209	data, 0.35	data, 0.275	hl, 0.446	data, 0.326	data, 0.241	data, 0.291	ohdsi, 0.293
covid, 0.209	cooky, 0.313	belrai, 0.188	arwen, 0.212	citizen, 0.269	fhir, 0.205	eu, 0.176	health, 0.121	spaces, 0.19	cooky, 0.229
esfri, 0.195	fiware, 0.21	twinning, 0.144	hospital, 0.166	tehdas, 0.237	ihe, 0.168	open, 0.165	research, 0.073	space, 0.133	pioneer, 0.176
european strategy forum on research infrastructures (esfri)	fiware (foundation)	belrai	arwen	healthy data e-consultation	helios	big data value public-private partnership (bdv)	actionable rwe network (arwen)	big data value association (bdva)	european health data and evidence network (ehden) (incl. ehden academy)
opensafely	health outcomes observatory (h2o)	dhu (digital health uptake) and dhu radar	athena (augmenting therapeutic effectiveness through novel analytics)	tehdas (towards european health data space)	hl7 (health level seven) international	data.europa.eu	amdex	data space support centre (dssc)	ohdsi (observational health data sciences and informatics)
phiri (population health information research infrastructure)	nihr health informatics collaborative	digitalhealtheurope catalogue of digital solutions supporting the digital transformation of health and care			hl7 belgium	dataeuropa	bdva	vlaamse smart data space (vsds)	pioneer
		gaia-x			integrating the healthcare enterprise (ihe)	european open science cloud (eosc)	big data for better outcomes (bd4bo)		
						european platform on rare disease registration	darwin (data analysis and real-world interrogation network in the european union)		
							data saves lives		
							data sharing coalition		