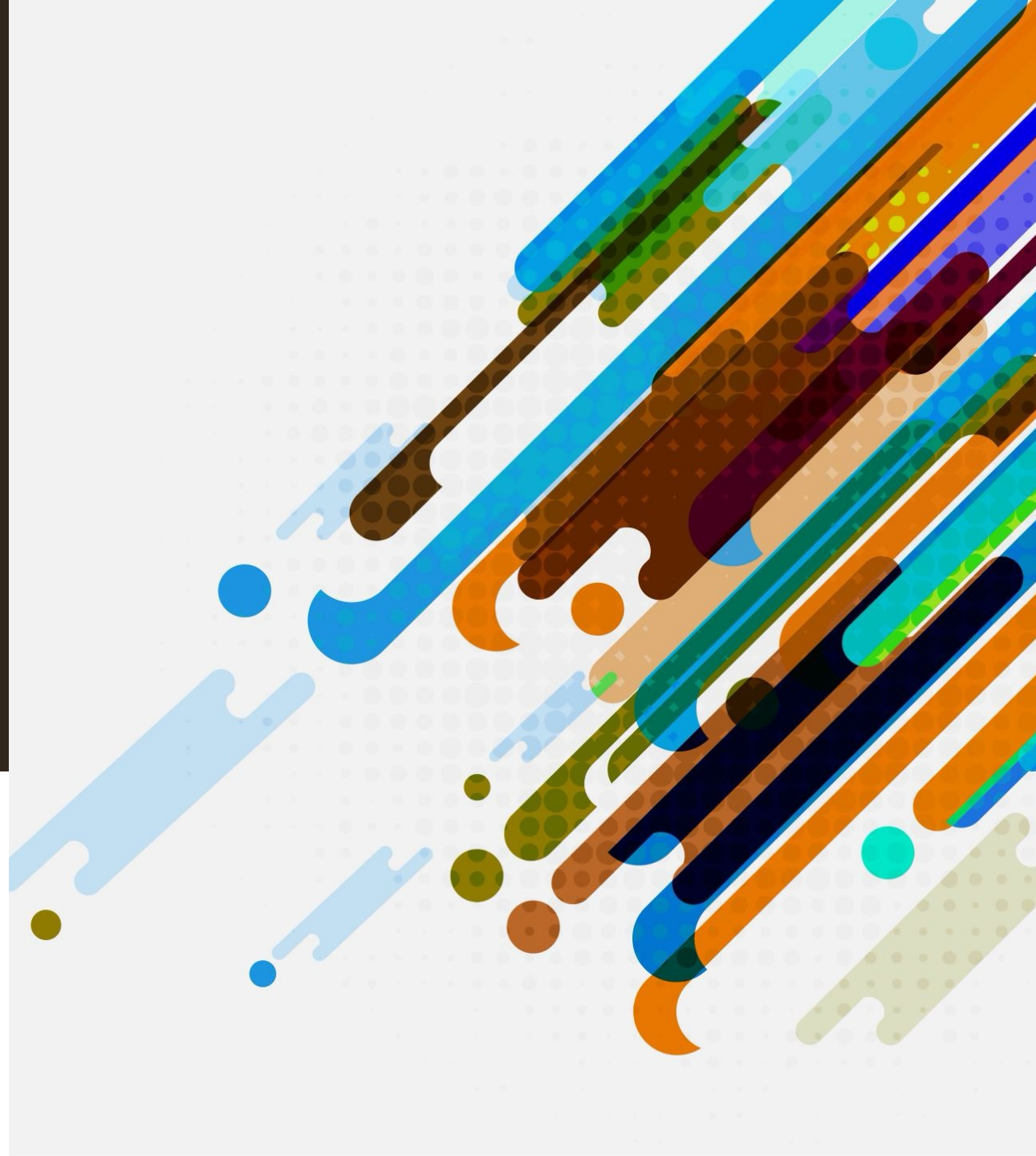
Topic modelling for text documents using NLP techniques

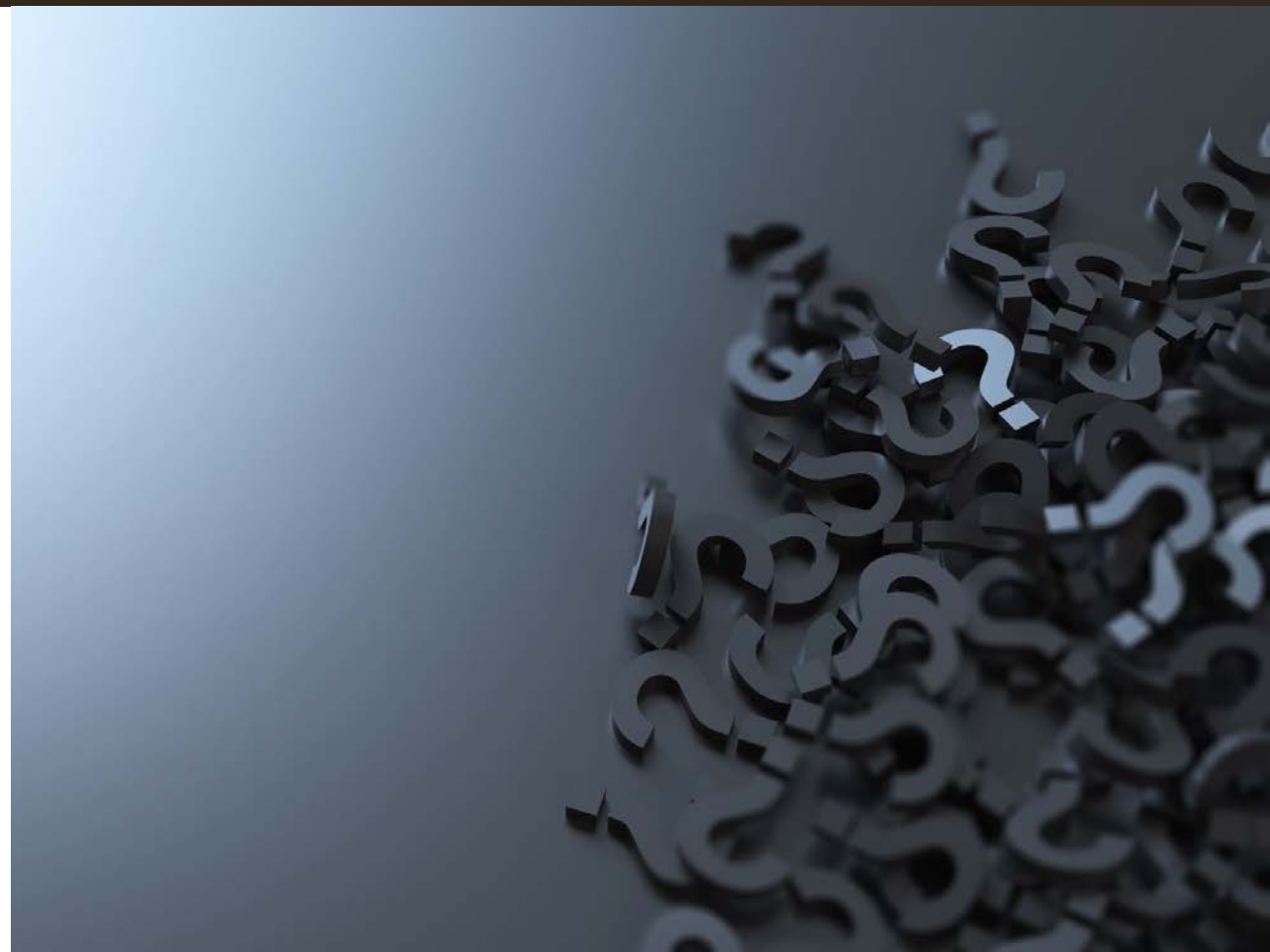
Akshay Jain

Engineering Manager



Use Case

- Identify entities within documents
- Data cleansing and document similarity
- Build documents clusters
- Build hierarchy across topics to query data based on sub sectors within documents



Identify entity documents

Andy Murray chalked up a 500th hard-court win in Dubai to join an exclusive club containing greats Roger Federer, Novak Djokovic, Andre Agassi and Rafael Nadal - then sparked more speculation over his future by suggesting he only has a few months left of his career.

Fresh from a chastening defeat to 18-year-old Jakub Mensik in the Qatar Open last week, Murray produced a superb comeback to beat Denis Shapovalov on Monday.

It was Murray's biggest result of the season - albeit a season that has yielded just two wins from six tournaments.

Identify entities within documents

```
import spacy
import pandas as pd

nlp = spacy.load("en_core_web_lg")
doc = nlp(text)

ent_label = []
[ent_label.append([ent.text, ent.label_]) for ent in doc.ents]

df = pd.DataFrame(ent_label, columns=['entity', 'entity_type']).drop_duplicates()
print(df)
```

Identify entities within documents

	entity	entity_type
0	Andy Murray	PERSON
1	500th	ORDINAL
2	Dubai	GPE
3	Roger Federer	PERSON
4	Novak Djokovic	PERSON
5	Andre Agassi	PERSON
6	Rafael Nadal	PERSON
7	a few months	DATE
8	18-year-old	DATE
9	Jakub Mensik	PERSON
10	the Qatar Open	EVENT
11	last week	DATE
12	Murray	PERSON
13	Denis Shapovalov	PERSON
14	Monday	DATE
16	the season albeit a season	DATE
17	just two	CARDINAL
18	six	CARDINAL

Identify entities within documents

Entity Type	Description
PERSON	People, including fictional ones.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non - GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LAW	Named documents made into laws.
LANGUAGE	Any named language.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage, including "%".
MONEY	Monetary values, including unit.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	"first", "second", etc.
CARDINAL	Numerals that do not fall under another type.

Identify similar documents

```
import spacy

nlp = spacy.load("en_core_web_lg")  # make sure to use larger package!
doc1 = nlp("I like salty fries and hamburgers.")
doc2 = nlp("Fast food tastes very good.")
# Similarity of two documents
print(doc1, "<->", doc2, doc1.similarity(doc2))


# Similarity of tokens and spans
french_fries = doc1[2:4]
burgers = doc1[5]
print(french_fries, "<->", burgers, french_fries.similarity(burgers))
```

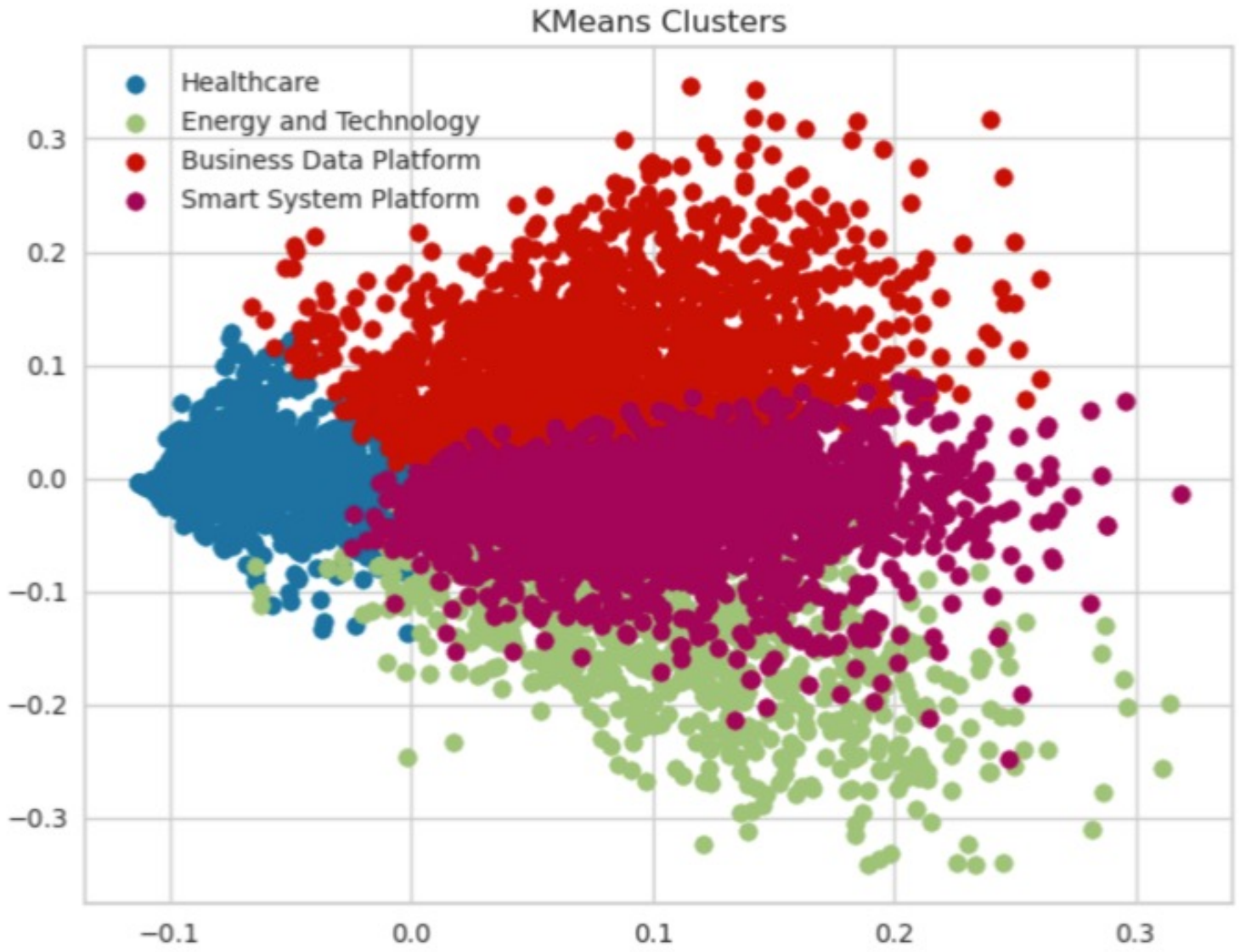
[53] ✓ 0.8s

```
I like salty fries and hamburgers. <-> Fast food tastes very good. 0.6871286202797843
salty fries <-> hamburgers 0.6901010870933533
```

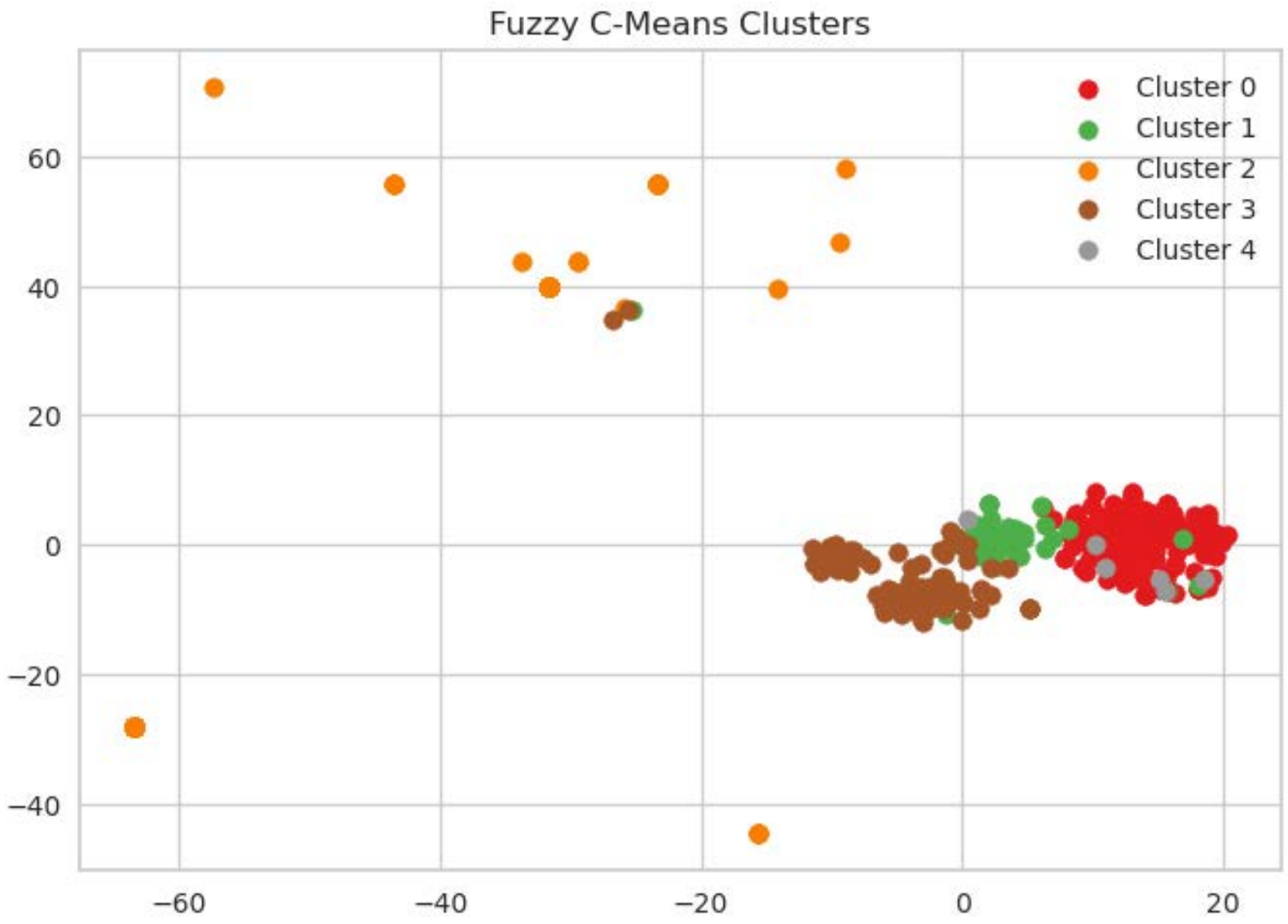
Clean data

```python
def preprocess_text(text):
    # Step 1: Lowercase the text
    text = text.lower()

    # Step 2: Remove punctuation
    text = ''.join([char for char in text if char not in string.punctuation])

    # Step 3: Tokenization
    tokens = nltk.word_tokenize(text)

    # Step 4: Remove stopwords
    stop_words = set(stopwords.words('english'))
    tokens = [word for word in tokens if word not in stop_words]

    # Step 5: Lemmatization
    lemmatizer = WordNetLemmatizer()
    tokens = [lemmatizer.lemmatize(word) for word in tokens]

    # Step 6: Remove specific words
    words_to_remove = ["project", "aim", "scope", "team", "objective", "key"]
    tokens = [word for word in tokens if word not in words_to_remove]

    # Step 7: Apply additional pre-processing steps
    text = text.replace(',', '')
    text = text.rstrip('\n')
    text = text.casefold()
    text = re.sub('\W_', ' ', text)
    text = re.sub("\S*\d\S*", " ", text)
    text = re.sub("\S*@\S*\s?", " ", text)

    return ' '.join(tokens)
```

Build documents clusters



Build documents clusters



Build documents clusters

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
0	health social care	greenhouse gas emission	main area focus	coffee cup lid
1	main area focus	main area focus	natural language processing	main area focus
2	artificial intelligence ai	uk supply chain	new revenue stream	dental practice dental
3	improve patient outcome	reduce carbon footprint	artificial intelligence ai	olawuyi racett nigeria
4	mental health problem	hydrogen fuel cell	machine learning algorithm	rock bolt tester
5	mental health issue	net zero carbon	ai machine learning	malignant brain tumour
6	machine learning algorithm	single use plastic	minimum viable product	pet coffee cup
7	mental health support	sustainable development goal	social medium platform	bovine tuberculosis btb
8	nh foundation trust	reduce carbon emission	year5 postproject revenue	communication dental practice
9	physical mental health	electric vehicle ev	sustainable development goal	practice dental laboratory

Cluster 0: Healthcare

Cluster 1: Energy and Technology

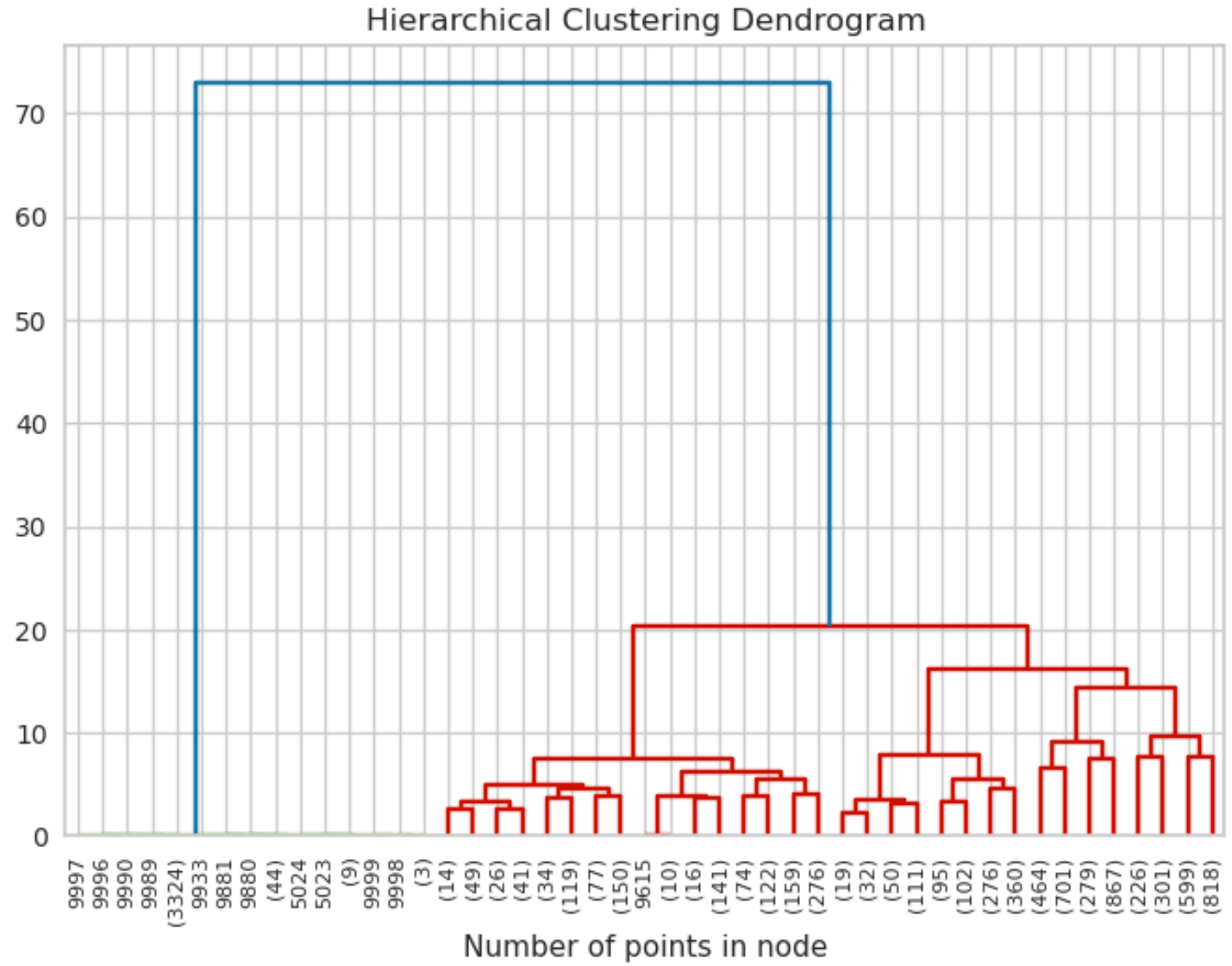
Cluster 2: Business Data Platform

Cluster 3: Smart System Platform

Build documents clusters

	Bigrams	Trigrams
0	(environmental impact, 123)	(government food strategy, 49)
1	(soil health, 108)	(greenhouse gas emission, 42)
2	(food security, 107)	(sustainable resilient farming, 36)
3	(net zero, 97)	(integrated pest management, 33)
4	(climate change, 88)	(soil organic matter, 31)
5	(abiotic stress, 66)	(net zero emission, 29)
6	(food production, 66)	(progression net zero, 25)
7	(crop yield, 64)	(environmental impact farming, 24)
8	(greenhouse gas, 64)	(sustainability environmental impact, 23)
9	(food strategy, 61)	(biotic abiotic stress, 21)
10	(sustainable resilient, 56)	(dissemination knowledge exchange, 19)
11	(supply chain, 56)	(longer term resilience, 16)
12	(farming practice, 56)	(precision breeding act, 16)
13	(resilient farming, 55)	(reduce environmental impact, 15)
14	(knowledge exchange, 55)	(genetic technology precision, 15)
15	(government food, 51)	(technology precision breeding, 15)
16	(ghg emission, 51)	(breeding act 2023, 15)
17	(pest management, 51)	(english sugar beet, 15)
18	(soil resilience, 50)	(benefit societal impact, 14)
19	(gas emission, 45)	(antimicrobial resistance amr, 13)

Agglomerative Clustering



Email: akshay.data.engineer@gmail.com

Linkedin: www.linkedin.com/in/akshaydjain