Untitled

April 27, 2020

Jacob Duvall

1 the_summarizer

[70]:

```
import os
import random
import glob
from sklearn.metrics import silhouette_score
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.cluster import KMeans
import json
import pickle
import nltk
import numpy as np
import pandas as pd
import networkx
import os


# look at the format of the json schema
def look_at_the_format_of_the_file(file):
    file = open(file, 'r')
    print(file.read())


# Takes glob and randomly extracts a percent of the files in a list
def choose_documents(folders, percent):
    file_list = list()
    for folder in folders:
        files = os.listdir(folder)
        for file in files:
            file_list.append(file)
    size = int(len(file_list) * (percent*(10**-2)))
    return random.sample(file_list, size)
```

```
# Takes a list of files and tokenizes the data in the files using␣
 ↪CountVectorizer
def files_reader(files_list):
    directory = 'C:\\Users\\jdale\\OneDrive\\School\\Text Analytics\\' \
                '*\\*\\pdf_json\\'
    single_large_document_list = list()
    for file in files_list:
        dir = glob.glob(str(directory + file))
        with open(dir[0]) as file_json:
            data = json.load(file_json)
            document_string = ''
            for line in data['body_text']:
                document_string = document_string + line['text']
            single_large_document_list.append(document_string)

    return single_large_document_list


# use tfidfvectorizer on the documents to get the matrix
def tfid_vectorize(doc):
    tv = TfidfVectorizer(min_df = 0., max_df=1., use_idf=True)
    tv_matrix = tv.fit_transform(doc)
    pickle.dump(tv_matrix, open('yummy_pickle.pkl', 'wb'))
    return tv, tv_matrix


# use countvectoizer on the documents to get the matrix
def count_vectorize(doc):
    cv = CountVectorizer()
    cv_matrix = cv.fit_transform(doc)
    pickle.dump(cv_matrix, open('yummy_pickle_cv.pkl', 'wb'))
    return cv, cv_matrix


# Takes tokenized documents and clusters them -- Uses Silhouette Coefficient to␣
 ↪measure cluster quality.
# Records the documents that are part of each cluster
def cluster_documents(doc_list):
    cv, tokenized_files = count_vectorize(doc_list)
    cluster_range = list(range(2, 8))
    for n_clusters in cluster_range:
        km = KMeans(n_clusters=n_clusters)
        km_predicts = km.fit_predict(tokenized_files)
        pickle_save = 'pickle_km_' + str(n_clusters) + '.pkl'
        pickle.dump(km_predicts, open(pickle_save, 'wb'))

        score = silhouette_score(tokenized_files, km_predicts)
        print("Number of clusters: {}, Silhouette Score: {}".format(n_clusters,␣
 ↪score))
    return cv, tokenized_files


# fits the best cluster size and shows the feature names
def best_cluster(cv, tokenized_files, size):
    km = KMeans(n_clusters=size).fit(tokenized_files)
    feature_names = cv.get_feature_names()
    ordered_centroids = km.cluster_centers_.argsort()[:, ::-1]
    for cluster_num in range(2):
        print('CLUSTER #' +str(cluster_num+1))
        feature_list = list()
        for i in ordered_centroids[cluster_num, :10]:
            feature_list.append(feature_names[i])
        print(feature_list)


# summarize the documents looking at their top 8 sentences through TextRank
def summarize_documents(string_list):
    count = 0
    for string in string_list:
        sentences = nltk.sent_tokenize(string)
        try:
            if len(sentences) < 8:
                raise Exception
        except Exception:
            continue
        tv, tv_matrix = tfid_vectorize(sentences)
        try:
            tv_matrix = tv_matrix.toarray()
        except:
            print('too large')
            continue
        similarity_matrix = np.matmul(tv_matrix, tv_matrix.T)
        try:
            similarity_graph = networkx.from_numpy_array(similarity_matrix)
        except MemoryError:
            print('memory error 1')
            continue
        try:
            scores = networkx.pagerank(similarity_graph)
        except MemoryError:
            print('memory error')
            continue
        ranked_sentences = sorted(((score, index) for index, score in scores.items()), reverse=True)
        top_sentenc_indices = [ranked_sentences[index][1] for index in range(8)]

        top_sentenc_indices.sort()
        write_summary_to_file(count, np.array(sentences)[top_sentenc_indices])
        count += 1


# write the summaries generated from summarize_documents() to file
def write_summary_to_file(count, summary_array):
    #print(count)
    if count < 1:
        try:
            os.remove("SUMMARY.md")
        except:
            pass
        file = open("SUMMARY.md", "w",  encoding='utf-8')
        header = 'This file was generated using TextRank summarization. The process taken was to extract ' \
                 'a random percent sampling of pdf_json files from ' \
                 'https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge. With these ' \
                 'files extracted, I then opened all files, tokenized the sentences of the ' \
                 'files using tfid vectorizer, ' \
                 'and then applied a TextRank algorithm to the tokenized docs. This TextRank allowed me to then ' \
                 'extract the 8 most useful summarized sentences. And voila!'
        file.write(header)
        file.write('\n\n')
        file.close()

    file = open("SUMMARY.md", "a",  encoding='utf-8')
    file.write(str(summary_array))
    file.write('\n\n')
    file.close()
```

2 main

2.1 Look at the format of the file

[39]:

```
look_at_the_format_of_the_file('C:\\Users\\jdale\\OneDrive\\School\\Text Analytics\\json_schema.txt')
```

```
# JSON schema of full text documents
```

```
{
    "paper_id": <str>,                      # 40-character sha1 of the PDF
    "metadata": {
        "title": <str>,
        "authors": [                        # list of author dicts, in order
            {
                "first": <str>,
                "middle": <list of str>,
                "last": <str>,
                "suffix": <str>,
                "affiliation": <dict>,
                "email": <str>
            },
            ...
        ],
        "abstract": [                       # list of paragraphs in the abstract
            {
                "text": <str>,
                "cite_spans": [             # list of character indices of
inline citations
                                            # e.g. citation "[7]" occurs at
positions 151-154 in "text"
                                            #      linked to bibliography entry
BIBREF3
                    {
                        "start": 151,
                        "end": 154,
                        "text": "[7]",
                        "ref_id": "BIBREF3"
                    },
                    ...
                ],
                "ref_spans": <list of dicts similar to cite_spans>,     # e.g.
inline reference to "Table 1"
                "section": "Abstract"
            },
            ...
        ],
        "body_text": [                      # list of paragraphs in full body
                                            # paragraph dicts look the same as
above
            {
                "text": <str>,
                "cite_spans": [],
                "ref_spans": [],
                "eq_spans": [],
                "section": "Introduction"
            },
            ...
            {
                ...,
                "section": "Conclusion"
            }
        ],
        "bib_entries": {
            "BIBREF0": {
                "ref_id": <str>,
                "title": <str>,
                "authors": <list of dict>       # same structure as earlier,
                                                # but without `affiliation` or
`email`
                "year": <int>,
                "venue": <str>,
                "volume": <str>,
                "issn": <str>,
                "pages": <str>,
                "other_ids": {
                    "DOI": [
                        <str>
                    ]
                }
            },
            "BIBREF1": {},
            ...
            "BIBREF25": {}
        },
        "ref_entries":
            "FIGREF0": {
                "text": <str>,                  # figure caption text
                "type": "figure"
            },
            ...
            "TABREF13": {
                "text": <str>,                  # table caption text
                "type": "table"
            }
        },
        "back_matter": <list of dict>           # same structure as body_text
    }
}
```

2.2 Choose documents

[45]:

```
sampling_of_files = choose_documents(glob.glob(
    'C:\\Users\\jdale\\OneDrive\\School\\Text Analytics\\*\\*\\pdf_json'), 10)
import pprint
pprint.pprint(sampling_of_files[:10])
```

```
['4c159330655513984fac5ab8e0c575512d33b514.json',
 'eb265b07935811bc052549f3780c9af843ecf45e.json',
 '8da224b5eba13cee28c3a95cb3c1e3c4d26e5a3a.json',
 '5e5908b9f7ad23dd332a32e1dfa411c4b0e88f59.json',
 'ddb7090314e89cfc001186101093a15896ba3b43.json',
 'b142d0d3130f98fe1bc2cec14d62de4362c80534.json',
 'd3d1b82a318250a2e7193878c11a4e14817b07eb.json',
 '49b80f692b4d3e2ebbd59454e10c29d522d8c2f7.json',
 '9f5625d9287d8c6729215a20add99c8cd4ba08cb.json',
 '0953fa36903063f60627e07f7b4e07f0aec3c4d3.json']
```

2.3 Write a files reader

[48]:

```
string_of_files_list = files_reader(sampling_of_files)
pprint.pprint(string_of_files_list[:2])
```

```
['The global assurance of safe drinking water and basic sanitation has been '
 'recognised as a United Nations Millennium Development Goal 1 , particularly '
 'in light of the pressures of rising urbanisation, agricultural '
 'intensification and climate change 2,3 . These trends enforce an increasing '
 'demand for freshwater monitoring frameworks that combine cost effectiveness, '
 'fast technology deployability and data transparency 4 . Environmental '
 'metagenomics, the tracing of organisms present in a substrate through '
 'high-throughput DNA sequencing, yields informative measures of relative '
 'taxonomic species occurrence and functional diversity 5 . Microbial '
 'metagenomics studies overcome enrichment biases common to traditional '
 'culturing approaches 5 ; however, they usually depend on expensive and '
 'stationary equipment, highly specialised operational training and '
 'substantial time lags between fieldwork, sample preparation, raw data '
 'generation and access.In recent years, these challenges have been revisited '
 "with the prospect of 'portable' DNA analysis. The main driver of this is the "
 'smartphone-sized MinION device from Oxford Nanopore Technologies (ONT), '
 'which enables real-time DNA sequencing using nanopores 6 . Nanopore read '
 'lengths can be comparably long (currently up to ~2*10 6 bases 7 ), which is '
 'enabled by continuous electrical sensing of sequential nucleotides along '
 'single DNA strands. In connection with a laptop or cloud access for the '
 'translation of raw voltage signal into nucleotides, nanopore sequencing can '
 'be used to rapidly monitor long DNA sequences in remote locations. Although '
 "there are still common concerns about the technology's base-level accuracy, "
 'mobile MinION setups have already proven powerful for real-time tracing and '
 'open data sharing during bacterial and viral pathogen outbreaks [8] [9] [10] '
```

```
'[11] [12] [13] .Here we report a simple, inexpensive workflow to assess '
'microbial freshwater ecosystems with nanopore DNA sequencing. Our benchmark '
'involves the design and optimisation of essential experimental steps for '
'multiplexed MinION usage in the context of local environments, together with '
'an evaluation of computational methods for the bacterial classification of '
'nanopore sequencing reads from metagenomic libraries. To showcase the '
'resolution of sequencing-based aquatic monitoring in a spatiotemporal '
'setting, we combine DNA analyses with physicochemical measurements of '
'surface water samples collected at nine locations within a confined ~12 '
'kilometre reach of the River Cam passing through the city of Cambridge (UK) '
'in April, June and August 2018.Nanopore full-length (V1-V9) 16S ribosomal '
'RNA (rRNA) gene sequencing was performed on all locationbarcoded freshwater '
'samples at each of the three time points (Figure 1 ; Supplementary Table 1a '
'). Samples were complemented with a negative control (deionised water) and a '
'mock community control composed of eight bacterial species in known mixture '
'proportions (Methods).To obtain valid taxonomic assignments from freshwater '
'sequencing profiles using nanopore sequencing, twelve different '
'classification tools were compared through several performance metrics '
'(Extended Data Figure 1 ; Methods). Root mean square errors (RMSE) between '
'observed and expected bacteria of the mock community differed slightly '
'across all classifiers. An Enterobacteriaceae overrepresentation was '
'observed across all replicates and classification methods, pointing towards '
'a consistent Escherichia coli amplification bias potentially caused by '
'skewed taxonomic specificities of the selected 16S primer pair (27f and '
'1492r) 14 . Robust quantifications were obtained by Minimap2 15 alignments '
'against the SILVA v.132 database 16 , for which 99.68 % of classified reads '
'aligned to the expected mock community taxa (mean sequencing accuracy 92.08 '
'%). Minimap2 classifications reached the second lowest RMSE (excluding '
'Enterobacteriaceae), and relative quantifications were highly consistent '
'between mock community replicates. Benchmarking of the classification tools '
"on one aquatic sample further confirmed Minimap2's reliable performance in a "
'more complex bacterial community, although other tools such as SPINGO 17 , '
'MAPseq 18 , or IDTAXA 19 also produced highly concordant results despite '
'variations in processing speed and memory usage (data not shown).Using '
'Minimap2 classifications within our bioinformatics consensus workflow '
'(Extended Data Figure 2 ;Methods), we then inspected sequencing profiles of '
'three independent MinION runs for a total of 30 river DNA isolates and six '
'controls. This yielded ~8. 3 and Rhodobacteraceae (2.5 %) (Figure 2d ). '
'Members of these families are commonly associated with aquatic environments; '
'for example, Burkholderiaceae reads mostly originate from genera such as '
'Limnohabitans, Rhodoferax or Aquabacterium, which validates the suitability '
'of this nanopore metagenomics workflow.Hierarchical clustering additionally '
'showed that two biological replicates collected at the same location and '
'time point (April samples 9.1 and 9.2), grouped with high concordance; this '
'indicates that spatiotemporal trends are discernible even within a highly '
'localised context. Besides the dominant core microbiome, microbial profiles '
'showed a marked arrangement of time dependence, with water samples from '
'April grouping more distantly to those from June and August (Figure 2c ). '
```

```
'Principal component analysis (PCA) (Figure 3a ; Extended Data Figure   4 ) '
'revealed that the strongest differential abundances along the chronological '
'axis of variation (PC3) derived from the higher abundance of '
'Carnobacteriaceae in April (Figure 3b ). This family is known for its '
'occurrence in waters with low temperature 20 .While a seasonal difference in '
'bacterial composition can be expected due to increasing water temperatures '
'in the summer months, additional changes may have also been caused by '
'alterations in river hydrochemistry and flow rate (Extended Data Figures 5 '
'and 6 , respectively; Supplementary Table 1c) . To assess this effect in '
'detail, we measured the pH and a range of major and trace cations in all '
'river water samples using inductively coupled plasma-optical emission '
'spectroscopy (ICP-OES), as well as major anions using ion chromatography '
'(Extended Data Figure 5 ; Methods). As with the bacterial composition '
'dynamics, we observed significant temporal variation in water chemistry, '
'superimposed on a spatial gradient of generally increasing sodium and '
'chloride concentrations along the river reach. This spatially consistent '
'effect is likely attributed to wastewater and agricultural discharge inputs '
'in and around Cambridge city. A comparison of the major element chemistry in '
"the River Cam transect with the world's 60 largest rivers further "
'corroborates the likely impact of anthropogenic pollution in this fluvial '
'ecosystem 21 (Extended Data Figure 5 ; Methods).In line with these '
'physicochemical trends, we next determined the spatiotemporal enrichment of '
'potentially functionally important bacterial taxa through nanopore '
'sequencing. We retrieved 55 potentially pathogenic bacterial genera through '
'careful integration of species known to affect human health 22, 23 , and '
'also 13 wastewaterassociated 24 bacterial genera (Supplementary Table 3 ). '
'Of these, 21 potentially pathogenic and eight wastewaterassociated genera '
'were detected across all of the river samples (Figure 3c ; Methods). Many of '
'these signals were stronger downstream of urban sections, within the mooring '
'zone for recreational and residential barges (location 7, Figure 1a ) and in '
'the vicinity of sewage outflow from a nearby wastewater treatment plant '
'(location 8). The most prolific candidate pathogen genus observed was '
'Arcobacter, which features multiple species implicated in acute '
'gastrointestinal infections 25 .In general, much of the taxonomic variation '
'across all samples was caused by sample April-7 (PC1 explains 27. Using '
'multiple sequence alignments between nanopore reads and pathogenic species '
'references, we further resolved the phylogenies of three common potentially '
'pathogenic genera occurring in our river samples, Pseudomonas, Legionella '
'and Salmonella (Extended Data Figure 7 ; Methods). While Legionella and '
'Salmonella diversities only presented negligible levels of known harmful '
'species, a cluster of sequencing reads in downstream sections indicated a '
'low abundance of the opportunistic, environmental pathogen Pseudomonas '
'aeruginosa (Extended Data Figure 7 ).We also found significant variations in '
'relative abundances of the Leptospira genus, which was recently described to '
'be enriched in wastewater effluents in Germany 27 . Indeed, the peak of '
'River Cam Leptospira reads falls into an area of increased sewage influx '
'(Figure 3c ). The Leptospira genus contains several potentially pathogenic '
'species capable of causing life-threatening leptospirosis through waterborne '
```

```
'infections 28 , however also features closerelated saprophytic and '
"'intermediate' taxa 29 . To resolve its complex phylogeny in the River Cam "
'surface, we aligned Leptospira reads from all samples together with various '
'reference sequences assigned to pre-classified pathogenic, saprophytic and '
'other environmental Leptospira species 29 (Figure 3d ; Supplementary Table '
'4a; Methods). Despite the presence of nanopore sequencing errors (Extended '
'Data Figure 8 ) and correspondingly inflated read divergence, we could '
'pinpoint spatial clusters and a distinctly higher similarity between our '
'amplicons and saprophytic rather than pathogenic Leptospira species. These '
'findings were subsequently validated by targeted, Leptospira '
'species-specific qPCR (Supplementary Table 5 , Methods), confirming that the '
'latest nanopore sequencing quality is high enough to yield indicative '
'results for bacterial monitoring workflows at the species level.Using an '
'inexpensive, easily adaptable and scalable framework, we provide the first '
'spatiotemporal nanopore sequencing atlas of bacterial microbiota along a '
'river reach. Beyond the core microbiome of an example fluvial ecosystem, our '
'results suggest that it is possible to robustly assess the heterogeneity in '
'accessory bacterial composition in the context of supporting physical '
'(temperature, flow rate) and hydrochemical (pH, inorganic solutes) '
"parameters. We show that the technology's current accuracy of ~92 % allows "
'for the designation of significant human pathogen community shifts along '
'rural-to-urban river transitions, as illustrated by downstream increases in '
'the abundance of pathogen candidates.Furthermore, our assessment of popular '
'bioinformatics workflows for taxonomic classification highlights current '
'challenges with error-prone nanopore sequences. We observed differences in '
'terms of bacterial quantifications, read misclassification rates and '
'consensus agreements between the twelve tested computational methods. In '
'this computational benchmark, using the SILVA v.132 reference database, one '
'of the most balanced performances was achieved by Minimap2 alignments. As '
'nanopore sequencing quality continues to increase through refined pore '
'chemistries, basecalling algorithms and consensus sequencing workflows [30] '
'[31] [32] , future bacterial taxonomic classifications are likely to improve '
'and advance opportunities for aquatic species discovery.We show that '
'nanopore amplicon sequencing data can resolve the core microbiome of a '
'freshwater body, as well as its temporal and spatial fluctuations. Besides '
'common freshwater bacteria, we find that the differential abundances of '
'Carnobacteriaceae most strongly contribute to seasonal loadings in the River '
'Cam.Carnobacteriaceae have been previously associated with cold environments '
'20 , and we found these to be more abundant in colder April samples (mean '
'11.3 řC, vs. 15.8 řC in June and 19.1 řC in August). This might help to '
'establish this family as an indicator for bacterial community shifts along '
'with water temperature fluctuations.Most routine freshwater surveillance '
'frameworks focus on semi-quantitative diagnostics of only a limited number '
'of target taxa, such as pathogenic Salmonella, Legionella and faecal '
'coliforms 33, 34 . Our proof-of-principle analysis highlights that the '
'combination of full-length 16S rRNA gene amplification and nanopore '
'sequencing can complement hydrochemical controls in pinpointing relatively '
'contaminated freshwater sites, some of which had been previously highlighted '
```

```
'for their pathogen diversity and abundance of antimicrobial resistance genes '
'35, 36 .Nanopore sequencing here allowed for the reliable distinction of '
'closely related pathogenic and non-pathogenic bacterial species of the '
'common Salmonella, Legionella, Pseudomonas and Leptospira genera. For '
'Leptospira bacteria, we further validated nanopore sequencing results '
'through the gold standard qPCR workflow of Public Health England '
'(Supplementary Table 5 ).A number of experimental intricacies should be '
'addressed towards future nanopore freshwater sequencing studies, mostly by '
'scrutinising water DNA extraction yields, PCR biases and molar imbalances in '
'barcode multiplexing ( Figure 2a ; Extended Data Figure 8 ). Yet, our '
'results show that it would be theoretically feasible to obtain meaningful '
'river microbiota from >100 barcoded samples on a single nanopore flow cell, '
'thereby enabling water monitoring projects involving large collections at '
'costs below č20 per sample (Supplementary Table 6 ). Barcoded shotgun '
'nanopore sequencing protocols may pose a viable alternative strategy to '
'bypass pitfalls often observed in amplicon-based workflows, namely '
'taxon-specific primer biases 14 Competing interests: All authors of this '
'manuscript declare no competing interest.Stammnitz (maxrupsta@gmail.com), or '
'to Andre Holzer (andre.holzer.biotech@gmail.com) and Lara Urban '
'(lara.h.urban@gmail.com). CC-BY-ND 4.0 International license author/funder. '
'It is made available under a The copyright holder for this preprint (which '
'was not peer-reviewed) is the . https://doi.org/10.1101/2020.02.06.936302 '
'doi: bioRxiv preprint The copyright holder for this preprint (which was not '
'peer-reviewed) is the . https://doi.org/10.1101/2020.02.06.936302 doi: '
'bioRxiv preprint The copyright holder for this preprint (which was not '
'peer-reviewed) is the . https://doi.org/10.1101/2020.02.06.936302 doi: '
'bioRxiv preprint The copyright holder for this preprint (which was not '
'peer-reviewed) is the . https://doi.org/10.1101/2020.02.06.936302 doi: '
'bioRxiv preprint Q1 -1.5*IQR (lower), and Q3 + 1.5*IQR (upper), '
'respectively; Q1: first quartile, Q3: third quartile, IQR: interquartile '
'range. The centre colour-scale table depicts the categorisation of subsets '
'of genera as waterborne bacterial pathogens (WB), drinking water pathogens '
'(DWP), potential drinking water pathogens (pDWP), human pathogens (HP) and '
'core genera from wastewater treatment plants (WW) (dark grey: included, '
'light grey: Table 3 ). The right-hand circle plot shows the distribution of '
'bacterial genera across locations of the River Cam. Circle sizes represent '
'overall read size fractions, while circle colours (sigma scheme) represent '
'the standard deviation from the observed mean relative abundance within each '
'genus. (d) Phylogenetic tree illustrating the multiple sequence alignment of '
'all nanopore reads classified as Leptospira, together with known Leptospira '
'reference sequences ranging from pathogenic to saprophytic species 29 '
'(Supplementary Table   4a ). The copyright holder for this preprint (which '
'was not peer-reviewed) is the . https://doi.org/10.1101/2020.02.06.936302 '
'doi: bioRxiv preprint MATERIALS AND METHODSWe monitored nine distinct '
'locations along a 11.62 km reach of the River Cam, featuring sites upstream, '
'downstream and within the urban belt of the city of Cambridge, UK. '
'Measurements were taken at three time points, in two-month intervals between '
'April and August 2018 (Figure 1; Supplementary Table 1a ). To warrant river '
```

```
'base flow conditions and minimise rain-derived biases, a minimum dry weather '
'time span of 48h was maintained prior to sampling 51 . One litre of surface '
'water was collected in autoclaved DURAN bottles (Thermo Fisher Scientific, '
'Waltham, MA, USA), and cooled to 4 řC within three hours. Two bottles of '
'water were collected consecutively for each time point, serving as '
'biological replicates of location 9 (samples 9.1 and 9.2).We assessed '
'various chemical, geological and physical properties of the River Cam '
'(Extended Data Figures 5 and   6 , Supplementary Tables 1b and 1c) .In situ '
'water temperature was measured immediately after sampling. To this end, we '
'linked a DS18B20 digital temperature sensor to a portable custom-built, grid '
'mounted Arduino nano v3. The copyright holder for this preprint (which was '
'not peer-reviewed) is the . https://doi.org/10.1101/2020.02.06.936302 doi: '
'bioRxiv preprint indicated precision of more than 4 % for Cl -. However, the '
'high Clconcentrations of the samples in this study were not fully bracketed '
'by the calibration curve and we therefore assigned a more conservative '
'uncertainty of 10 % to Clconcentrations.High calcium and magnesium '
'concentrations were recorded across all samples, in line with hard '
'groundwater and natural weathering of the Cretaceous limestone bedrock '
'underlying the river catchment (Extended Data Figure 5 ).There are no known '
'evaporite salt deposits in the river catchment, and therefore the high '
'dissolved Na + , K + and Clconcentrations in the River Cam are likely '
'derived from anthropogenic inputs 52 (Extended Data Figure 5 ). We The total '
'dissolved solid (TDS) concentration across the 30 freshwater samples had a '
'mean of 458 mg/L (range 325 -605 mg/L) which is relatively high compared to '
'most rivers, due to 1.) substantial solute load in the Chalk groundwater '
'(particularly Ca 2+ , Mg 2+ , and HCO3 -) and 2.) likely anthropogenic '
'contamination (particularly Na + , Cl -, and SO4 2-). The TDS range and the '
'major ion signature of the River Cam is similar to other anthropogenically '
'heavily-impacted rivers 21 , exhibiting enrichment in Na + (Extended Data '
'Figure 5 ).Overall, ion profiles clustered substantially between the three '
'time points, indicating characteristic temporal shifts in water chemistry. '
'PC1 of a PCA on the solute concentrations [ţmol/L] shows a strong time '
'effect, separating spring (April) from summer (June, August) samples '
'(Extended Data Figure 5b) . We highlighted the ten most important features '
'(i.e., features with the largest weights) and their contributions to PC1 '
'(Extended Data Figure   5c ).We integrated sensor data sets on mean daily '
'air temperature, sunshine hours and total rainfall from a public, The '
'copyright holder for this preprint (which was not peer-reviewed) is the . '
'https://doi.org/10.1101/2020.02.06.936302 doi: bioRxiv preprintWithin 24 '
'hours of sampling, 400 mL of refrigerated freshwater from each site was '
'filtered through an individual 0.22 ţm pore-sized nitrocellulose filter '
'(MilliporeSigma, Burlington, MA, USA) placed on a Nalgene polysulfone bottle '
'top filtration holder (Thermo Fisher Scientific) at -30 mbar vacuum '
'pressure. Additionally, 400 mL deionised (DI) water was also filtered. We '
'then performed DNA extractions with a modified DNeasy PowerWater protocol '
'(Qiagen, Hilden, Germany). Briefly, filters were cut into small slices with '
'sterile scissors and transferred to 2 mL Eppendorf tubes containing lysis '
'beads. Homogenization buffer PW1 was added, and the tubes subjected to ten '
```

```
'minutes of vigorous shaking at 30 Hz in a TissueLyser II machine (Qiagen). '
'After subsequent DNA binding and washing steps in accordance with the '
"manufacturer's protocol, elution was done in 50 ţL EB. We used Qubit dsDNA "
'HS Assay (Thermo Fisher Scientific) to determine water DNA isolate '
'concentrations (Supplementary Table 2a ).DNA extracts from each sampling '
'batch and DI water control were separately amplified with V1-V9 full-length '
'Figure 1b-c) , which was previously assessed using nanopore shotgun '
'metagenomics 42 . We used common primer binding sequences 27f and 1492r, '
'both coupled to unique 24 bp barcodes and a nanopore motor protein tether '
'sequence (Supplementary Table 7 ). Fulllength 16S rDNA PCRs were performed '
'with the following conditions: Table S2b ). We used KAPA Pure Beads (KAPA '
'Biosystems, Wilmington, MA, USA) to concentrate full-length 16S rDNA '
'products in 21 ţL DI water. Multiplexed nanopore ligation sequencing '
'libraries were then made by following the SQK-LSK109 protocol (Oxford '
'Nanopore Technologies, Oxford, UK).R9.4 MinION flow cells (Oxford Nanopore '
'Technologies) were loaded with 75 ţl of ligation library. The MinION '
'instrument was run for approximately 48 hours, until no further sequencing '
'reads could be collected. Fast5 files were basecalled using Guppy (version '
'3.15) and output DNA sequence reads with Q>7 were saved as fastq '
'files.Various output metrics per library and barcode are summarised in '
'Supplementary Table 2c .In collaboration with Public Health England, raw '
'water DNA isolates of the River Cam from each location and time point were '
'subjected to the UK reference service for leptospiral testing (Supplementary '
'Table 5 ). This test is based on quantitative real-time PCR (qPCR) of 16S '
'rDNA and LipL32, implemented as a TaqMan assay for the detection and '
'differentiation of pathogenic and non-pathogenic Leptospira spp. from human '
'serum. Briefly, the assay consists of a two-component PCR; the first '
'component is a duplex assay that targets the gene encoding the outer '
'membrane lipoprotein LipL32, which is reported to be strongly associated '
'with the pathogenic phenotype.The second reaction is a triplex assay '
'targeting a well conserved region within the 16S rRNA gene (rrn) in The '
'described data processing and read classification steps were implemented '
'using the Snakemake workflow management system 53 and are available on '
'Github -together with all necessary downstream analysis scripts to reproduce '
'the results of this manuscript (https://github.com/d-j-k/puntseq). We '
'gathered read statistics such as quality scores and read lengths using '
'NanoStat (version 1.1.2, https://github.com/wdecoster/nanostat), and used '
'Pistis (https://github.com/mbhall88/pistis) to create quality control plots. '
'This allowed us to assess GC content and Phred quality score distributions, '
'which appeared consistent across and within our reads. Overall, we obtained '
'2,080,266 reads for April, 737,164 for June, and 5,491,510 for August, with '
'a mean read quality of 10.0 (Supplementary Table 2c ).We used twelve '
'different computational tools for bacterial full-length 16S rDNA sequencing '
'read classification We used nanopore sequencing data from our mock community '
'and freshwater amplicons for benchmarking the classification tools. We '
'therefore subsampled (a) 10,000 reads from each of the three mock community '
'sequencing replicates (section 1.4), and (b) 10,000 reads from an aquatic '
'sample (April-8; three random draws served as replicates). We then used the '
```

```
'above twelve classification tools to classify these reads against the same '
'database,SILVA v.132 16 (Extended Data Figure 1 ).For the mock community '
'classification benchmark, we assessed the number of unclassified reads, '
'misclassified reads (i.e. sequences not assigned to any of the seven '
'bacterial families), and the root mean squared error (RMSE) between observed '
'and expected taxon abundance of the seven bacterial families. Following the '
'detection of a strong bias towards the Enterobacteriaceae family across all '
'classification tools, we also analysed RMSE values after exclusion of this '
'family (Extended Data Figure 1b -c).For the aquatic sample, the number of '
'unclassified reads were counted prior to monitoring the performance of each '
'classification tool in comparison with a consensus classification, which we '
'defined as majority vote across classifications from all computational '
'workflows. We observed stable results across all three draws of 10,000 reads '
'from the same dataset (data not shown), indicating a robust representation '
'of the performance of each classifier.Minimap2 performed second best at '
'classifying the mock community (lowest RMSE), while also delivering '
'freshwater bacterial profiles in line with the majority vote of other '
'classification tools (Extended Data Figure 1d e), in addition to providing '
'rapid speed (data not shown). Yet, the application of this software to our '
'entire dataset caused insufficient memory errors (at ~150 Gb RAM with kmer '
'length 12), likely due to major sequence redundancies within the SILVA v.132 '
'reference fasta file. Therefore, to run each of our full samples within a '
'reasonable memory limit of 50 Gb, it was necessary to reduce the number of '
"threads to 1, raise the kmer size ('k') to 15 and set the minibatch size "
"('-K') to 25M (i.e., the number of query bases that are processed at any "
'time), prolonging the runtime of several samples to ~three days.After '
'applying Minimap2 to the processed reads as explained above (section 2.2.4), '
'we processed the resulting SAM files by firstly excluding all header rows '
"starting with the '@' sign and then transforming the sets of read IDs, SILVA "
'IDs, and alignment scores to TSV files of unique read-bacteria assignments '
'either on the bacterial genus or family level. All reads that could not be '
'assigned to the genus or family level were discarded, respectively.In the '
'case of a read assignment to multiple taxa with the same alignment score, we '
'determined the lowest taxonomic level in which these multiple taxa would be '
'included. If this level was above the genus or family level, respectively, '
'we discarded the read.Across three independent sequencing replicates of the '
'same linear bacterial community standard (section 2.2.1),we found that the '
'fraction of reads assigned to unexpected genus level taxa lies at ~1 % when '
'using the Minimap2 classifier and the SILVA v.132 database.Raw quantified '
'DNA, PCR amplicons and sequencing read counts were considerably less '
'abundant in DI water negative controls, as compared to actual freshwater '
'specimens (Supplementary Table 2a Table 8 ).Minimap2 alignments against mock '
'community taxa were used to determine the mean read-wise nanopore sequencing '
'accuracy for this study, as determined by the formula: accuracy = 1 -(read '
'mismatch length œ read alignment length)These values were calculated for '
'each of all eight species against each sequencing replicate, using the '
'samtools 66 (v.1.3.1) stats function.Sample-specific rarefaction curves were '
'generated by successive subsampling of sequencing reads classified by Figure '
```

```
'3 ; section 2.4.1). Although we mainly present a single example rarefied '
'dataset within this manuscript, we repeated each analysis, including PCAs, '
'hierarchical clustering and Mantel tests, based on additional rarefied '
'datasets to assess the stability of all results.We performed Mantel tests '
'(using scikit-bio version 0.5.1) to compare rarefied datasets with the full '
'dataset. We All classification assessment steps and summary statistics were '
'performed in R or python (https://github.com/dj-k/puntseq). We used the '
"python package 'scikit-bio' for the calculation of the Simpson index and the "
"Shannon's diversity as well as equitability index.. CC-BY-ND 4.0 "
'International license author/funder. It is made available under a The '
'copyright holder for this preprint (which was not peer-reviewed) is the . '
'https://doi.org/10.1101/2020.02.06.936302 doi: bioRxiv preprintRarefied read '
'count data was subjected to a log10(x+1) transformation before hierarchical '
'clustering using the complete linkage method. For PCA analyses, rarefied '
'read count data was subjected to log10(x+1) and Ztransformations. Negative '
'control samples were removed. Mock community samples were initially removed '
'to then be re-aligned to the eigenspace determined by the aquatic samples. '
'We provide PCA visualisations of the main principal components (PCs '
'explaining >10 % variance, respectively). For each of these relevant PCs, we '
'further highlight the ten most important features (i.e., taxa with the '
'largest weights) and their contributions to the PCs in barplots.For '
'detecting outlier bacterial families per sample, we chose bacteria which '
'were 1.) identified by more than 500 reads and 2.) which were at least five '
'times more abundant in any single sample than in the mean of all samples '
'combined.A list of 55 known bacterial pathogenic genera, spanning 37 '
'families, was compiled for targeted sequence testing.This was done through '
'the careful integration of curated databases and online sources, foremost '
'using PATRIC 22 and data on known waterborne pathogens 23 (Supplementary '
'Table 3a ). Additionally, we integrated known genera from a large wastewater '
'reference collection 24 (Supplementary Table 3b ).To identify if DNA reads '
'assigned to Leptospiraceae were more similar to sequence reads of previously '
'identified pathogenic, intermediate or environmental Leptospira species, we '
'built a neighbour-joining tree of Leptospiraceae reads classified in our '
'samples data, together with sequences from reference databases (Figure 3d ; '
'species names and NCBI accession numbers in clockwise rotation around the '
'tree in Supplementary Table 4a ). We matched the orientation of our reads, '
'and then aligned them with 68 Leptospira reference sequences and the '
'Leptonema illini reference sequence (DSM 21528 strain 3055) as outgroup. We '
'then built a neighbour-joining tree using Muscle v.3.8.31 67 (excluding '
"three reads in the 'Other Environmental' clade that had extreme branch "
'lengths >0.2). The reference sequences were annotated as pathogenic and '
'saprophytic clades P1, P2, S1, S2 as recently described 29 .Additional '
'published river water Leptospira that did not fall within these clades were '
"included as 'Other Environmental' 68 . Similarly, we constructed phylogenies "
'for the Legionella, Salmonella and Pseudomonas genus, using established '
'full-length 16S reference species sequences from NCBI (Supplementary Table '
'4b-d).This study was designed to enable freshwater microbiome monitoring in '
'budget-constrained research environments. Although we had access to basic '
```

```
'infrastructure such as pipettes, a PCR and TissueLyser II machine, as well a '
'high-performance laptop, we wish to highlight that the total sequencing '
'consumable costs were held below č4,000 (Supplementary Table 6a ). Here, '
'individual costs ranged at ~č75 per sample (Supplementary Table   6b ). With '
'the current MinION flow cell price of č720, we estimate that per-sample '
'costs could be further reduced to as low as ~č15 when barcoding and pooling '
'~č100 samples in the same sequencing run (Supplementary Table   6c ). '
'Assuming near-equimolar amplicon pooling, flow cells with an output of '
'~5,000,000 reads can yield well over 37,000 sequences per sample and thereby '
'surpass this conservative threshold applied here for comparative river '
'microbiota analyses.Sequencing datasets generated and analysed during this '
'study are available from the European Nucleotide The are no restrictions on '
'data availability.Our Github repository (https://github.com/d-j-k/puntseq/) '
'provides a Snakemake framework that integrates all data pre-processing '
'steps, and a Singularity that contains all necessary software '
'(https://github.com/d-jk/puntseq/tree/master/analysis/). We further provide '
'complete and rarefied SILVA 132 classifications from runs of Minimap2 '
'(https://github.com/d-j-k/puntseq/tree/master/minimap2_classifications), '
'which can be directly used as an input for downstream analyses.. CC-BY-ND '
'4.0 International license author/funder. It is made available under a The '
'copyright holder for this preprint (which was not peer-reviewed) is the . '
'https://doi.org/10.1101/2020.02.06.936302 doi: bioRxiv preprint (a-b) List '
'of pathogen (a) and wastewater (b) candidate bacterial genera. and August '
'(c).. CC-BY-ND 4.0 International license author/funder. It is made available '
'under a The copyright holder for this preprint (which was not peer-reviewed) '
'is the . https://doi.org/10.1101/2020.02.06.936302 doi: bioRxiv preprint',
'Coronavirus (CoV) remains a public health concern 16 years after the '
'outbreak of the severe acute respiratory syndrome coronavirus (SARS-CoV) in '
'2003 [1] . The Middle East respiratory syndrome coronavirus (MERS-CoV) '
'emerged in 2012, reemerged in 2015, and is still circulating in the Middle '
'East region, which reminds the international community that the threat of '
'CoVs persists [2, 3] . However, neither vaccine nor drugs against CoVs are '
'currently available. Outbreaks of CoVs initiated extensive structural '
'investigation on CoV encoded proteins thereafter, which not only shed light '
'on the life cycle of CoVs but also laid foundation for the structure-based '
'drug design (SBDD). CoV contains a positive single-stranded RNA genome of~30 '
'kb, one of the largest among +RNA viruses [4, 5] . To maintain the unusually '
'large RNA genome, CoV encodes two replicase polyproteins pp1a and pp1ab, '
'which are broken down into 16 nonstructural proteins (nsps) via proteinase '
'cleavage [6, 7] . The nsps are then recruited to cytoplasm membranes, on '
'which they form the membraneassociated replication-transcription complex '
'(RTC). An RNAdependent RNA polymerase nsp12 and a helicase nsp13 are the '
'central components of RTC [8, 9] . However, while high-resolution structures '
'of most CoV encoded proteins had been determined soon after SARS-CoV '
'outbreak, the first CoV nsp13 structure, MERS-CoV nsp13, was only solved '
'recently [10] . Nsp13 belongs to helicase superfamily 1 and shares conserved '
'features with the eukaryotic Upf1 helicase [11, 12] . Nsp13 is a '
'multi-domain protein comprising of an N-terminal Cys/His rich domain (CH '
'domain) and a C-terminal SF1 helicase core [10] . Nsp13 exhibits multiple '
'enzymatic activities, including hydrolysis of NTPs and dNTPs, unwinding of '
'DNA and RNA duplexes with 5 0 -3 0 directionality and the RNA 5 0 '
'-triphosphatase activity [13, 14] . To investigate the structure of CoV '
'nsp13, we overexpressed the full-length MERS-CoV nsp13 (1-598aa) in insect '
'cells and purified. The activity of the recombinant MERS-CoV nsp13 was '
'verified by ATPase and helicase assays. Crystallization of MERS-CoV nsp13 '
'was achieved by adding a synthetic single-stranded 15 poly dT DNA with 5 0 '
'-triphosphate (ppp-15 T) to the protein, which restrains the intrinsic '
'flexibility of nsp13. Benefiting from the presence of an N-terminal '
'zinc-binding domain with three zinc atoms, multi-wavelength anomalous '
'diffraction (MAD) data at the zinc absorption edge was collected, which '
'allowed the determination of the crystal structure of MERS-CoV nsp13 [10] '
'.Prepare all solutions using ultrapure water (prepared by purifying '
'deionized water, to attain a sensitivity of 18 M-cm at 25 C) and analytical '
'grade reagents. Prepare and store all reagents at room temperature (unless '
'indicated otherwise). Diligently follow all waste disposal regulations when '
'disposing waste materials. We do not add sodium azide to reagents. 2. The '
'forward primer (gaaattggatccgctgtcggttcatgc) and the reverse primer '
'(gaaattctcgagtcactggagcttgtaatt) of full-length nsp13. Primers stocks are '
'either supplied or diluted by molecular biology grade water to 100 M and '
'stored at À20 C.3. The pFastbac-1 baculovirus transfer vector is modified; 6 '
'Â Histidine-SUMO tag with a C terminal PreScission protease (PPase) site '
'coding sequence in the N terminal of open reading frame [15] .All procedures '
'should be carried out at room temperature unless otherwise '
'specified.Transposition in E. coli DH10 Bac 1. Amplify MERS-nsp13 '
'full-length by PCR method with BamHI and XhoI restriction sites at 5 0 and 3 '
'0 termini, respectively.2. The amplified MERS-nsp13 gene should be digested '
'by BamHI/XhoI at 37 C for 1 h. The pFast-bac-6ÂHistidine-SUMO plasmid should '
'also be digested by BamHI/XhoI at the same time.3. Digested nsp13 DNA should '
'be ligated with pFast-bac-6ÂHistidine-SUMO vector using the rapid DNA '
'ligation kit. 1. Prepare 50 mL high-5 cells in express-5 medium at a density '
'of 0.38 Â 10 6 cells/mL, and culture in a 300 mL cell conical flask. '
'Incubate the culture at 28 C with shaking at 120 rpm for 48 h, the density '
'of cells will grow to 1.5-2.5 Â 10 6 cells/mL (see Note 2).2. Add 1.5 mL '
'MERS-CoV nsp13 P2 or P3 virus into the culture, and incubate at 22 C with '
'shaking at 120 rpm for 44-60 h.3. Centrifuge the culture at 3000 Â g for 30 '
'min. Collect the cells pellet. 10. Prepare the Ni-NTA resin, and add the '
'resin into 2-3 empty Econo-Columns (5 mL 50% resin per column), wash and '
'balance the resin with 10 mL lysis and wash buffer(II) twice.11. Place the '
'columns at 4 C. Apply the clarified cell lysates supernatant to the balanced '
'Ni-NTA resin, and flow through the column by gravity.12. Wash the resin in '
'the column with 10 mL lysis and wash buffer (II) 3 times.13. Resuspend the '
'resin by 3.5 mL L lysis and wash buffer(II), and add 100 L PPase. Incubate '
'the resin at 4 C for 10-12 h.14. Apply the buffer to the column and let it '
'flow under gravity. Collect the flow through in a 50 mL tube.15. Add another '
'25 mL lysis and wash buffer(II) to the resin and flow through the column. '
'Also collect the flow through in the previous 50 mL tube. 16 . To remove the '
```

```
'PPase, add the flow through to another column which contains the NS4B resin. '
'Collect the flow through from the NS4B resin column.17. Apply the flow '
'through to an Amicon Ultra protein concentrator (30 kDa filter, 50 mL), '
'centrifuge at 2465 Â g at 4 C until the sample volume is concentrated to 1 '
'mL. 18 . Transfer the concentrated sample to a 1.5 mL tube and centrifuge at '
'17,949 Â g for 3 min to remove the aggregates and particulates.19. Load the '
'sample onto the superdex 200 column in the size exclude chromatography (SEC) '
'buffer using an Ä KTA-purify chromatography at 4 C.20. Analyze 8 L of each '
'peak fractions by SDS-PAGE (Fig. 2) . 21 . Collect the fractions that '
'contain the single band of MERS-CoV nsp13, mix the fractions, and '
'concentrate the mixture to a final density of 6-8 mg/mL. 22 . 50 L packaged '
'the protein sample, quickly freeze them by liquid nitrogen and store them at '
'À80 C. 5. Spot 1 L sample from the mixture on the thin-layer chromatography '
'cellulose TLC plates and resolve with running buffer for 20 min.6. Dry the '
'plates and press the plate onto phosphor screen for 2 h. Analyze the result '
'by storage phosphor screen and Typhoon Trio Variable Mode Imager (Fig. 3) . '
'5. Take 4 L sample from each reaction mixtures and load the samples onto '
'10% native PAGE gel.6. Run the native PAGE gel at 100 V for 40 min on ice.7. '
'Scan the gel (Fig. 4 ).Crystals of the unliganded MERS-CoV nsp13 diffracted '
'the X-rays poorly, >3.6 Å . The addition of 5 0 -triphosphate-15 dT DNA '
'(ppp-15T) greatly improves the resolution. 2. Mix 1 L sample with 1 L '
'reservoir buffer from the crystallization conditions screen kits, and '
'incubate at 18 C using the hanging-drop vapor-diffusion system.3. '
'Crystallize MERS-CoV nsp13 by mixing with the equal volume of reservoir '
'buffer containing 0.1 M Tris-HCl (pH 8.5), 1 M (NH4) 2 SO 4 , and 15% '
'glycerol. Crystals grow to their maximum in a week (Fig. 5 ).Crystal '
'Structure 1. Highly redundant multi-wavelength anomalous diffraction data '
'should be collected using the X-ray with wavelengths close to the absorption '
'edge of zinc. High energy remote wavelength should be 1.2810 Å , peak '
'wavelength: 1.2827 Å (two datasets were collected to improve the '
'redundancy), and inflection wavelength 1.2831 Å .2. Data processing and '
'reducing by XDS Package and Truncate software from CCP4. The crystals belong '
'to the space group P6 1 22, and contained two copies of nsp13 per asymmetric '
'unit. 3. An interpretable electron density map should be calculated using '
'SHARP/autoSHARP [20] .Coot [21] .5. Collect native data with highest '
'resolution (3.0 Å ) using the X-rays with the wavelength of 0.978 Å .6. '
'Higher resolution structure should be solved by molecular replacement using '
'the initial nsp13 structure as the searching model. 7. Manual model building '
'with the improved electron density map. While most part of nsp13 can be '
'located, the electron density of 1B subdomain is very weak, reflecting that '
'this part is highly flexible.8. Structure refinement to resolution limit of '
'3.0 Å using software PHENIX [22] .In the final model (Fig. 6) , 145-230aa '
'(the entire 1B domain) of molecule A are disordered, probably due to '
'mobility of 1B and the lack of crystal contacts, whereas in molecule B, 591 '
'out of 598 amino acids were located in the electron density maps (Fig. 7) . '
'Data collection and refinement statistics are summarized in Table 1 . 4 '
'Notes 1. When we prepare P1 virus in six-well plates, the medium in the '
'wells always evaporated. Sealing the gap of the plate by medical tape can '
```

```
 "reduce the evaporation of medium (don't seal the gap completely, leave a "
 'small gap to keep the ventilation). Having a water trough in incubator also '
 'can reduce the evaporation of the medium.2. The culture of insect cells '
 'sometimes was harassed by the contamination of bacteria or other microbes. '
 'To avoid the contamination, we treat the conical flasks not only by '
 'conventional autoclave sterilization, but also leave the 3 L conical flask '
 '(sealed by tinfoil) in the oven at 200 C for 3-5 h before using.3. To remove '
 'nucleic acids bound to nsp13, we used the lysis buffer containing high '
 'concentrate salt; this is a key step and improves the crystallization of '
 'nsp13 [10] . In practice, when sonicated in the buffer containing high '
 'concentrate salt, we found that the SUMO-tagged recombinant proteins lead '
 'the supernatant of the high-5 cell lysate to be turbid, which finally blocks '
 'the affinity columns. We have tried four concentrations of NaCl in lysis '
 'buffer, including 300 mM, 500 mM, 1 M, and 1.5 M. The first three '
 "concentrations of NaCl render the supernatant to be unable to use, we can't "
 "improve it by highspeed centrifugation (47,850 Â g), and it also can't be "
 'filtered by 0.45 m syringe filter. The last concentration, 1.5 M NaCl in '
 'lysis buffer, could generate a bit better supernatant of cell lysates than '
 'other three concentrations of salt. We centrifuge the supernatant twice, '
 'then can filter it by 0.45 m syringe filters (100 mL supernatant consumed '
 'about 8-10 filters). This clarified supernatant can flow through the '
 'affinity columns well.4. The results of helicase assay always face the '
 'contamination of background fluorescence. Keep the gel from contacting any '
 'items containing fluorescence in the lab, including fluorescent dyes, some '
 'plastic boxes, hand towel, and so on.']
```

2.4 Cluster documents

[53]:

```python
cv, tokenized_files = cluster_documents(string_of_files_list)
```

```
Number of clusters: 2, Silhouette Score: 0.9270759361330172
Number of clusters: 3, Silhouette Score: 0.9207116157180787
Number of clusters: 4, Silhouette Score: 0.4672383758046355
Number of clusters: 5, Silhouette Score: 0.45250612718556243
Number of clusters: 6, Silhouette Score: 0.4552221519281297
Number of clusters: 7, Silhouette Score: 0.35165373560844365
```

[54]:

```python
best_cluster(cv, tokenized_files, 2)
```

```
CLUSTER #1
['the', 'of', 'and', 'in', 'to', 'is', 'with', 'di', 'or', 'for']
CLUSTER #2
['the', 'of', 'and', 'in', 'to', 'with', 'for', 'is', 'that', 'was']
```

2.5 Summarize documents

2.6 Write summarys to file

[ ]:

```
summarize_documents(string_of_files_list)
```

[72]:

```
file = open('C:\\Users\\jdale\\OneDrive\\School\\Text␣
 ↪Analytics\\project_2\\SUMMARY.md', 'r')
with open("C:\\Users\\jdale\\OneDrive\\School\\Text␣
 ↪Analytics\\project_2\\SUMMARY.md") as f:
    head = [next(f) for i in range(20)]
for i in head:
    print(i)
```

```
This file was generated using TextRank summarization. The process taken was to
extract a random percent sampling of pdf_json files from
https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge. With
these files extracted, I then opened all files, tokenized the sentences of the
files using tfid vectorizer, and then applied a TextRank algorithm to the
tokenized docs. This TextRank allowed me to then extract the 8 most useful
summarized sentences. And voila!



['The protein expressed from the IBV S1 protein multi-epitope cassette and the
NDV NP protein were found at 35 and 53 kDa bands in P1 and P20 of rNDV-IBV-T/B
lysates, however, the parental virus strain LaSota only had the~53 kDa NP
protein band (Figure 2) , confirming that rNDV-IBV-T/B successfully expressed
the IBV S1 protein multi-epitope cassette in vitro and that the protein was
antigenic, as evidenced by its recognition on a Western blot by anti-IBV serum.'

 'To determine whether inserting the IBV S1 protein multi-epitope cassette into
the NDV LaSota genome affects the biological properties of the recombinant
virus, the growth characteristics and pathogenicity after 20 passages of rNDV-
IBV-T/B were evaluated through virus titration, MDT, and ICPI tests in vitro and
in vivo.'

 'To determine whether inserting the IBV S1 protein multi-epitope cassette into
the NDV LaSota genome affects the biological properties of the recombinant
virus, the growth characteristics and pathogenicity after 20 passages of rNDV-
IBV-T/B were evaluated through virus titration, MDT, and ICPI tests in vitro and
in vivo.'

 'The HA titers of LaSota and rNDV-IBV-T/B were the same (2 10 ).'

 'The VN antibodies induced by rNDV-IBV-T/B administration could effectively
neutralize heterologous IBV strain, indicating that the recombinant vaccine
could protect against heterologous IBV challenge.To evaluate the efficacy of the
```

multi-epitope vaccine in protecting against IBV and NDV, one-day-old SPF chicks were immunized with rNDV-IBV-T/B virus or PBS via the oculonasal route.'

'The protection of chickens vaccinated with the multi-epitope vaccine was significantly higher than that of the PBS and LaSota vaccine groups (p < 0.05).'

'The hemagglutination inhibition (HI) titer against NDV and virus neutralization (VN) antibody titer against IBV are expressed as mean log2 Âś standard deviation.To evaluate the efficacy of the multi-epitope vaccine in protecting against IBV and NDV, oneday-old SPF chicks were immunized with rNDV-IBV-T/B virus or PBS via the oculonasal route.'

'The protection of chickens vaccinated with the multi-epitope vaccine was significantly higher than that of the PBS and LaSota vaccine groups (p < 0.05).']

['Only mannosylated E glycoproteins (which are exposed at the surface of DV virions transmitted to humans by infected mosquitoes), and not E proteins with complex glycosylation (produced in mammalian cells), have been shown to interact with DC-SIGN-expressing cells (22) .DC-SIGN and L-SIGN are endocytic receptors and their cytoplasmic tails carry putative internalisation signals such as a dileucine (LL) motif (which is present in both DC-SIGN and L-SIGN) and a tri-acidic cluster that is believed to be involved in intracellular trafficking (Fig.'

'We recommend the two following reviews for an overview of the physiological importance of LSECs in viral infections of the liver (6,7) .In this chapter, we will provide general protocols to study the molecular interactions between viruses and DC-SIGN or L-SIGN and to investigate the functions of these two lectins in viral infection and transmission.'

'We next present protocols to produce soluble viral envelope proteins using a Semliki forest virus (SFV) vector (35-37) and to study the molecular basis of viral capture and internalization by DC-SIGN or L-SIGN.'

'Note that only HIV gp120 DMJ , which carries only mannosylated N-glycans, binds to DC-SIGN.Wild-type viruses or viral particles carrying the reporter genes firefly luciferase (Luc) or green fluorescent protein (GFP) can both be used to study DC-SIGN and L-SIGN-mediated cis-infection and trans-enhancement of target cell infection.'

'DC-SIGN-or L-SIGN-expressing cells and their parental counterpart (10 5 cells) are exposed to viral particles for 2 h at 37ÂřC at varying MOI in FSC-free medium supplemented with 1% penicillin/streptomycin, pH approx 7.5.'

'Viral transmission is quantified two or 3 d later.DC-SIGN-or L-SIGN-expressing cells and their parental counterpart (10 5 cells) are incubated with viral particles at a low MOI (insufficient to directly infect target cells) for 2 h at 37ÂřC and immediately co-cultured with target cells without washing.'

'The protocol is similar to that described under Subheading 3.4.2. or 3.4.3. except that co-culture with target cells is started several days after exposure of cells expressing DC-SIGN or L-SIGN to virus.'

'The advantage of production of viral particles is that cells can be infected at equal MOI and hence the quantity of protein produced is more reproducible than direct electroporation with pSFV-âenv RNA.']