

DS4D Group 14

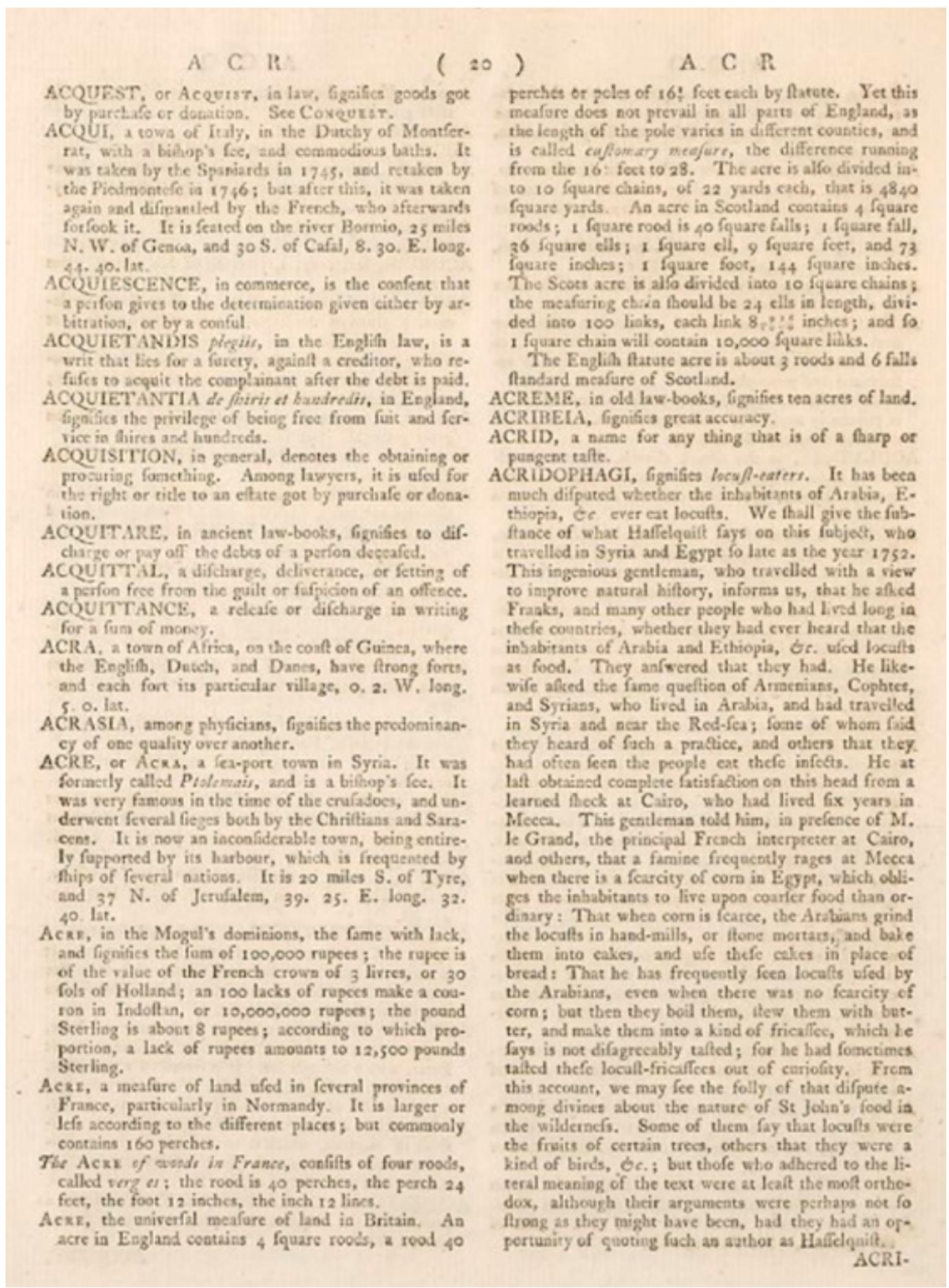
Encyclopedia Britannica

Dake Wang, Keren He, Qiuyue Ding, Shijie Chen, Yuhan Ma

Text data comes from OCR

Our text data comes from scanned copies of early encyclopedias. The earliest edition comes from 250 years ago (printing fades, missing), and there are 195 volumes in total. You can imagine how messy the results after scanning are

Original sample



Original sample

Edition 1	1771	Edition 5	1817
Edition 2	1784	Edition 6	1823
Edition 3	1797	Edition 7	1842
Edition 4	1810	Edition 8	1860

#1 Text Cleanup

1. Clean up each edition's entries

```
: # The first half of each chapter of the encyclopedia
word=""
word_dict={}
whole_sent=""

for sent in df_files_content["text"]:
    # If a sentence does not match the entry, it will
    res=re.match(r"[A-Z]+",sent)
    if res is None:
        whole_sent+=sent
    else:
        # If a sentence matches the entry, save the p
        word_dict[word]=whole_sent
        word=res.group().strip(",")
        whole_sent=sent.replace(word+",","")

# Add the last entry
word_dict[word]=whole_sent
```

A. Clean up entries

```
: # This is all the extracted entries
word_dict.keys()

: dict_keys(['SCO', 'AABAM', 'AADE',
ACATUAIA', 'ABACH', 'ABACO', 'ABACOT',
N', 'ABADIR', 'ABACRE', 'ABAFT', 'ABAI
A', 'ABARTICULATION', 'ABAS', 'ABAISED
EMENT', 'ABATIS', 'ABAVO', 'ABB', 'ABB
T', 'ABBREVIATION', 'ABBREVIATOR', 'ABI
RY', 'ABDALS', 'ABDELAVI', 'ABDEST', 'I
LIANS', 'ABELMOSCH', 'ABENSBURG', 'ABE
'ABERRATION', 'ABERYSWITH', 'ABESTA',
L', 'ABIB', 'ABIES', 'ABILITY', 'ABING
BLACTATION', 'ABLACQUEATION', 'ABLATIC
ION', 'ABO', 'ABOARD', 'ABOLITION', 'AI
RTION', 'ABORTIVE', 'ABOY', 'ABRA', 'AI
BRAUM', 'ABRASAX', 'ABRAX', 'ABREAST'.
```

B. All entries

```
df_encyclopedia_all.iloc[100:110,:]
```

	text	version
ABRA	a filver coin of Poland, in value nearly equivalent to an Englifli Hulling.	1
ABRACADABRA	a magical word or fspell, which being written as many times as the word contains letters, and omitting the lad letter of the former every time, was, in the ages of ignorance and fuperdition, vvyrn about the neck, as an antidote againt argues andfeveral other difeaies ABRAHAM'r balm, in botany, See Cannabis.	1
ABRAHAMITES	an order of monks exterminated - for idolatry by Theophilus in the ninth century. Alfo the name of another fedt of heretics who had adopted the errors of Paulus. See Paulicians.	1
ABRAMIS	an obfclete name for the filh cyprinus. See Cyprinus, B Abrasa, A B R (6 ABRASA, in furgery, ulcers, where the lkin is fo tender and lax as to render them fubjeft to abrafion.	1
ABRASION	in medicine, the corrodng of any part by acrid humours or medicines.	1
ABRAUM	an obfolete name of a certain fpecies of clay, called by fome authors Adamic earth, on account of its red colour.	1
ARRASAY	or Abraxas. a myfical term found in the ancient theology and philosophy of Balilides's followers	1

C. Collect all entries of each edition

```
: stop_words=stopwords.words('english')

def clean_text(text):

    # Remove all non-English letters directly
    pattern=r'[^a-z]'
    text=re.sub(pattern, ' ',text.lower())

    # Replace multiple consecutive spaces with 1 space
    text_list = re.sub(r'\s+', ' ', text).split()

    # Remove stop words
    text_list=[word for word in text_list if ((word not in stop_words) and len(word)>3)]
    return " ".join(text_list)
```

D. Keep cleaning

```
df_encyclopedia_all["text_clean"]=df_encyclopedia_all["text"].apply(clean_text)
df_encyclopedia_all.iloc[100:110,:]
```

	text	version	text_clean
ABRA	a filver coin of Poland, in value nearly equivalent to an Englifli Hulling.	1	filver coin poland value nearly equivalent englifli hulling
ABRACADABRA	a magical word or fspell, which being written as many times as the word contains letters, and omitting the lad letter of the former every time, was, in the ages of ignorance and fuperdition, vvyrn about the neck, as an antidote againt argues andfeveral other difeaies ABRAHAM'r balm, in botany, See Cannabis.	1	magical word fspell written many times word contains letters omitting letter former every time ages ignorance fuperdition vvyrn neck antidote againt argues andfeveral difeaies abraham balm botany cannabis
ABRAHAMITES	an order of monks exterminated - for idolatry by Theophilus in the ninth century. Alfo the name of another fedt of heretics who had adopted the errors of Paulus. See Paulicians.	1	order monks exterminated idolatry theophilus ninth century alfo name another fedt heretics

E. Output examples

#2 Using Topic Modeling and LDA to Analyze Encyclopedia

Wenbo Li, Le Sun, Yuanyong Feng, and Dakun Zhang. 2008. Smoothing LDA model for text categorization. In Proceedings of the 4th Asia information retrieval conference on Information retrieval technology (AIRS'08). Springer-Verlag, Berlin, Heidelberg, 83–94.

LDA model

```
from sklearn.feature_extraction.text import CountVectorizer
count = CountVectorizer(stop_words='english',
                        min_df=20,
                        max_df=0.1,
                        max_features=5000)
X = count.fit_transform(df_encyclopedia_all['text_clean'].values)

from sklearn.decomposition import LatentDirichletAllocation
lda = LatentDirichletAllocation(n_components=10,
                                 random_state=123)
X_topics = lda.fit_transform(X)
```

A.LDA model

Extract representative vocabulary for each topic

- Here are the top 20 important vocabularies for each topic

```
n_top_words = 20
feature_names = count.get_feature_names()

for topic_idx, topic in enumerate(lda.components_):
    print("Topic %d: " % (topic_idx + 1))
    print(" ".join([feature_names[i] for i in topic.argsort()[:n_top_words - 1]]))

Topic 1:
equal motion line point velocity force centre angle axis weight plane feet lines distance dia
meter parallel points circle inches greater
Topic 2:
muft fide fmall feet lefs furface diftance becaufe cafe fecond iron fquare earth piece placed
hand quantity half round inches
Topic 3:
small surface heat used iron acid glass light colour state matter quantity process substance
temperature produced white species size placed
Topic 4:
army emperor prince enemy troops kingdom empire battle obliged government rome arms duke roma
n peace romans power defeated military reign
Topic 5:
himself perfon faid againtf themfelves lord houfe perfons language mind prefent laft court th
```

B.20个主题中的重要的词汇

Topics in LDA model

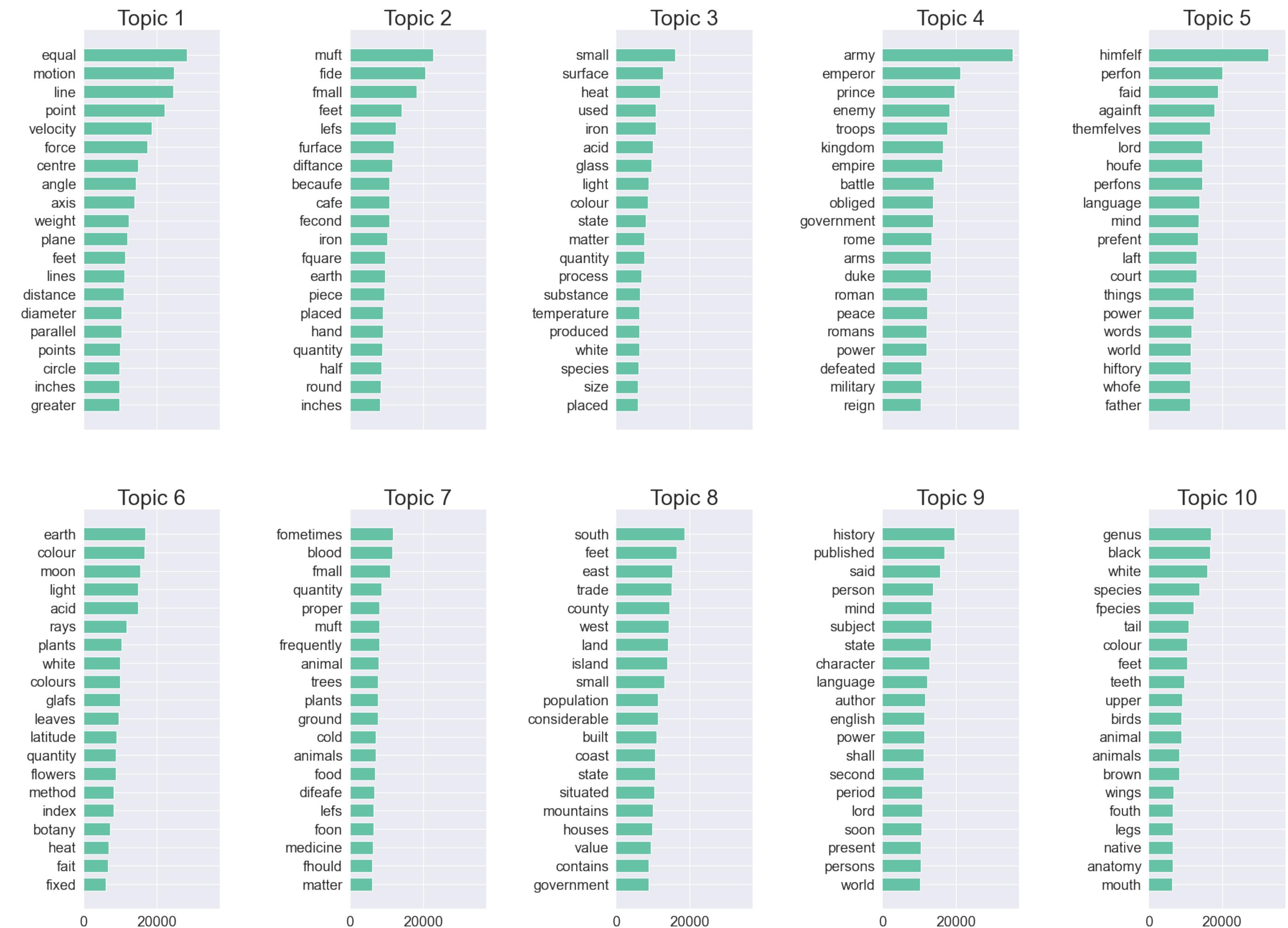
```

def plot_top_words(model, feature_names, n_top_words, title):
    fig, axes = plt.subplots(2, 5, figsize=(30, 20), sharex=True)
    axes = axes.flatten()
    for topic_idx, topic in enumerate(model.components_):
        top_features_ind = topic.argsort()[-n_top_words - 1:-1]
        top_features = [feature_names[i] for i in top_features_ind]
        weights = topic[top_features_ind]

        ax = axes[topic_idx]
        ax.barh(top_features, weights, height=0.7)
        ax.set_title(f'Topic {topic_idx + 1}', fontdict={'fontsize': 30})
        ax.invert_yaxis()
        ax.tick_params(axis='both', which='major', labelsize=20)
        for i in 'top right left'.split():
            ax.spines[i].set_visible(False)
        fig.suptitle(title, fontsize=40)

    plt.subplots_adjust(top=0.9, bottom=0.05, wspace=0.95, hspace=0.2)
    plt.show()

```



A.10 Topics and Their Words Frequency

#3 Name Each Topic

Topic 1:
equal motion line point velocity force centre angle axis weight plane feet lines distance diameter parallel points circle inches greater

Topic 2:
muft fide fmall feet lefs furface diftance becaufe cafe fecond iron fquare earth piece placed hand quantity half round inches

Topic 3:
small surface heat used iron acid glass light colour state matter quantity process substance temperature produced white species size placed

Topic 4:
army emperor prince enemy troops kingdom empire battle obliged government rome arms duke romana peace romans power defeated military reign

Topic 5:
himself perfon faid againtf themfelves lord houfe perfons language mind prefent laft court th ings power words world hifitory whofe father

Topic 6:
earth colour moon light acid rays plants white colours glafs leaves latitude quantity flowers method index botany heat fait fixed

Topic 7:
fometimes blood fmall quantity proper muft frequently animal trees plants ground cold animals food difeafe lefs foon medicine fhould matter

Topic 8:
south feet east trade county west land island small population considerable built coast state situated mountains houses value contains government

Topic 9:
history published said person mind subject state character language author english power shall second period lord soon present persons world

Topic 10:
genus black white species species tail colour feet teeth upper birds animal animals brown wings fouth legs native anatomy mouth

Topic 3: Materials science

Topic 4: Military & Royal Family

Topic 5: Politics & Church

Topic 7: Anatomy

Topic 8: Trade

Topic 9: Cultural

Topic1:

sample#1: PQ with a third plane FG, are parallel. For if the lines EF, GH, situated in the same plane, are not parallel, they must meet if produced ; therefore the planes MN, P Q, in which they are, must also meet, which is contrary to the hypothesis of their being parallel.

Topic2:

sample#1: the art of carting and working lead, and using it in building. As this metal melts soon and with little heat, it is easy to cart; it into figures of any kind, by running

Topic3:

sample#1: in building, a composition of white marble pulverised, and mixed with plaster of lime y and the whole being sifted and wrought np with water, is to be used like common plaster: this is called by Pliny marmoratum opus and albarivm opus. A s T U [787] S T U S

Topic4:

sample#1: a valiant Britifli queen in the time of Nero, the emperor, wife to Praefut

Topic5:

sample#1: (in Greekthe book'), a name applied by Chriftians, by way of eminence or distinction, to the collection of facred writings, or the holy fcriptures of the Old and New Testaments; known also by various other appellations, as, the Sacred Books, Holy Writ, Inspired Writings, Scriptures, &c. The Jews styled i the Bible (that is, the Old Testament) mikra; which signifies Lcijfon, or Lefture. This collection of the facred writings, containing those of the Old and New Testament, is justly looked ...

sample#2: a name applied by Chriftians, by way of eminence or distinction, to the collection of facred writings, or the holy fcriptures of the Old and New Testament; known also by various other appellations, as, the Sacred Books, Holy Writ, Inspired Writings, Scriptures, &c. The Jews styled the Bible (that is, the Old Testament) mikra, which signifies Lejfon, or Lefture. , This collection of the facred writings, containing those of the Old and New Testament, is judly looked upon as the foundation of the ...

sample#3: a general name given to the body of eccleſiaſtics of the Chriftian church, in contradistinction to the laity. See Laity. The distinction of Chriftians into clergy and

Topic 8 Trade

south, feet, east, **trade**, county, west, land, **island**, small, population, considerable, built, coast, state, situated, mountains, houses, **value**, contains, government, islands, british, extent, london, nearly, rivers, **bank**, **market**, **money**, house, chief, chiefly, building, produce, **corn**, soil, present, various, district, numerous, square, english, towns, extensive, western, northern, states, course, eastern, **cotton**

sample#1: a town situated on the isthmus of Panama. See Panama.PORTO FERRAJO, the capital of the Island of Elba, in the province of Pisa, and the duchy of Tuscany[^] celebrated as the residence of Napoleon during his banishment to that island. It is situated on a tongue of land running into the sea, and forming a small bay. It is fortified, and contains two churches, an hospital, 600 houses, with 3120 inhabitants, who depend chiefly upon some salt works and the tunny fishery. Lat. 42. 49. 6. Long. 9. 20. 1 ...

sample#2: or Newburgh, a borough-town of North Wales, in the island of Anglesey, and hundred of Menai, 257 miles from London, and twelve from Beaumaris. It was once the residence of the princes of Anglesey, and a corporation founded by Edward I. There is a market, which is held on Tuesday. The population amounted in 1801 to 599, in 1811 to 750, in 1821 to 756, and in 1831 to 804.NEW BRUNSWICK, a British province of North America, situated between the parallels of 45. 5. and 48. 4.30. north latitude, and ...

sample#3: or Carmarthenshire (Welsh Caerfyrddiri), a maritime county in South Wales, is bounded on the north by Cardigan, on the east by Brecon, on the south by Glamorgan and the Bristol Channel, and on the west by Pembroke. Its greatest length is from S.W. to N.E., about 52 miles; its greatest breadth, S.E. to N.W., about 28 miles. It possesses an area of 947 square miles, or 606,331 acres, and is thus the largest of all the Welsh counties. It contains 77 parishes, and is in the diocese of St David's. I ...

#4 Count the Distribution of entries for each topic by edition

In the same version, calculate the proportion of the number of entries for each topic

```
df_version_topic.div(df_version_topic.sum(axis=1),axis=0)
```

topic	0	1	2	3	4	5	6	7	8	9
version										
1	0.040850	0.105102	0.002563	0.054412	0.229968	0.084512	0.095923	0.039114	0.006037	0.341520
2	0.025595	0.123563	0.002205	0.071665	0.362262	0.061033	0.135769	0.031107	0.004725	0.182076
3	0.025297	0.118584	0.001715	0.072951	0.393667	0.074727	0.119258	0.031422	0.003430	0.158949
4	0.026834	0.122412	0.001877	0.072246	0.397198	0.092763	0.093388	0.036279	0.004128	0.152874

A. Percentage of entries

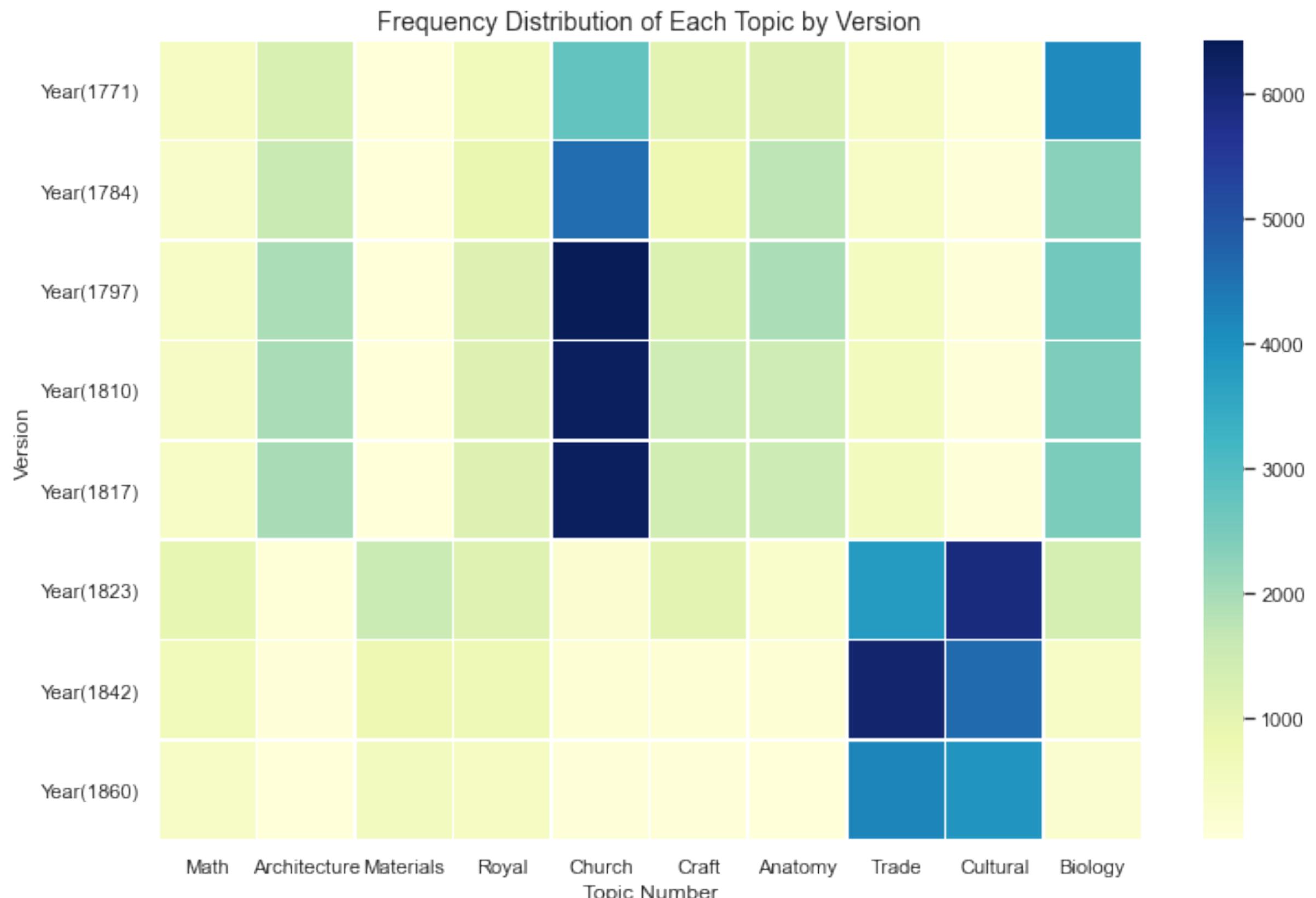
```
# Name the theme and visualize it
topic_labels=[ "Math", "Architecture", "Materials", "Royal", "Church", "Craft", "Anatomy", "T
plt.figure(figsize=(12,8))
sns.heatmap(df_version_topic, cmap="YlGnBu", linewidths=.5)

plt.xlabel("Topic Number", fontsize=12)
plt.xticks([x+0.5 for x in range(10)], topic_labels)

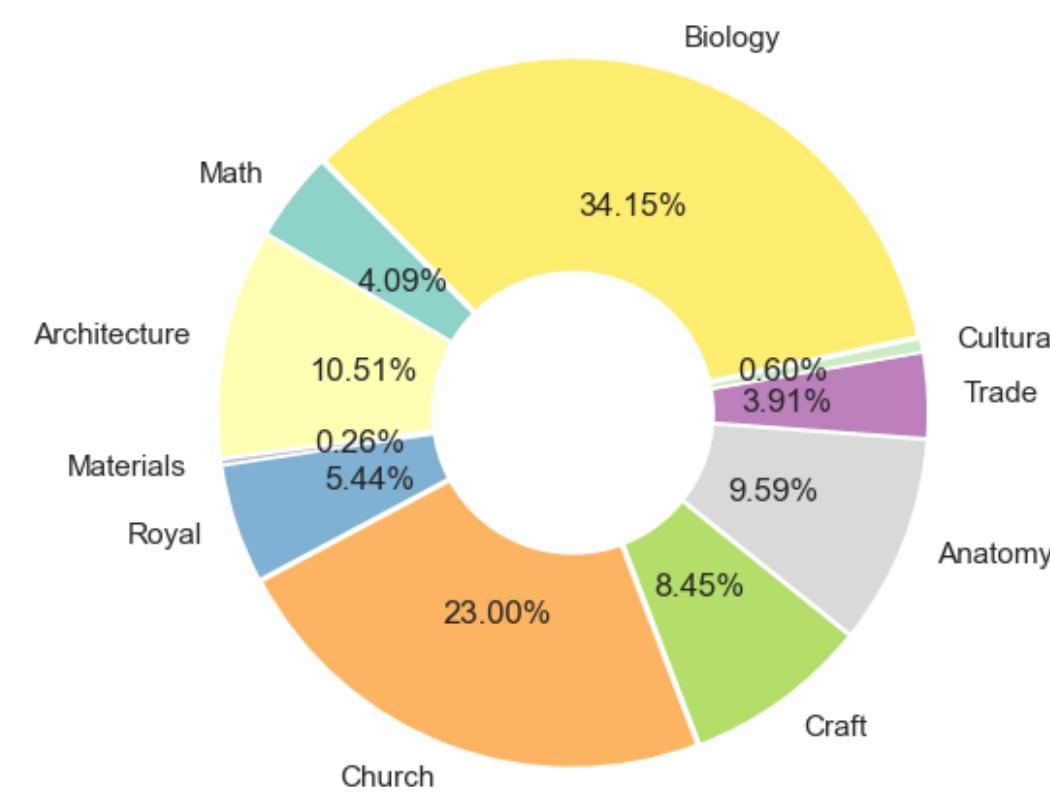
plt.ylabel("Version", fontsize=12)
plt.yticks([x+0.5 for x in range(8)],
           ['Year(1771)', 'Year(1784)', 'Year(1797)', 'Year(1810)', 'Year(1817)', 'Year(1820)', 'Year(1823)', 'Year(1826)'], rotation=0)

plt.title("Frequency Distribution of Each Topic by Version", fontsize=14)
```

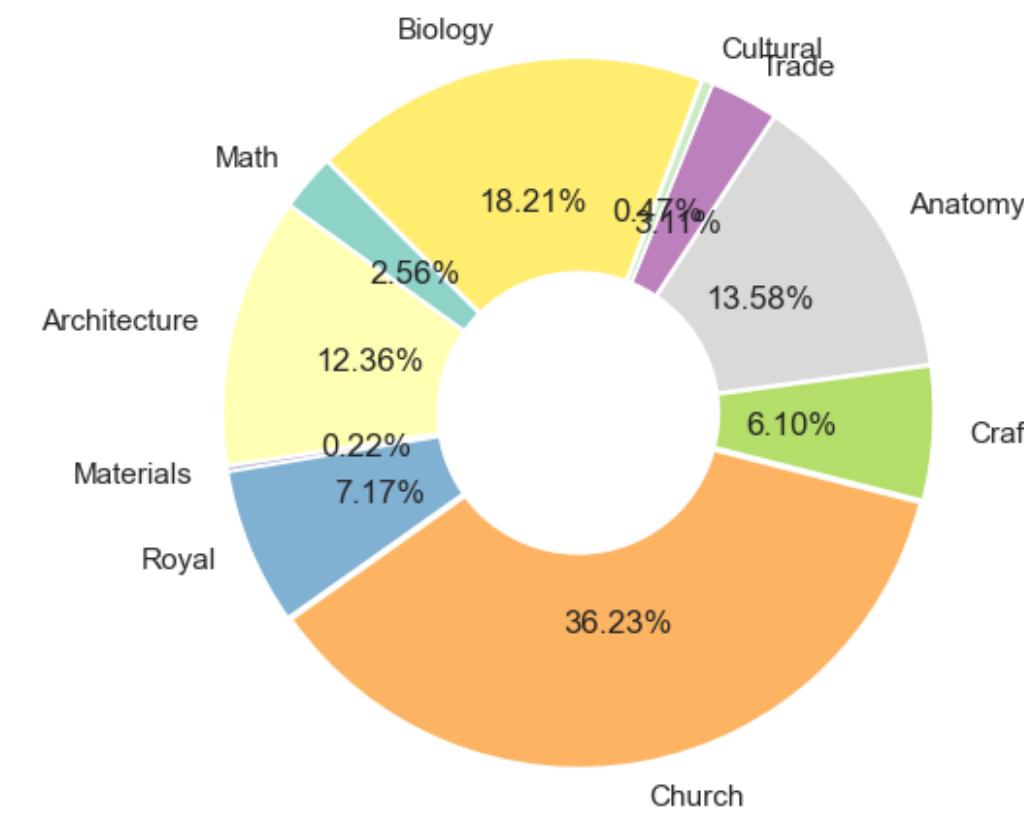
A. Percentage of Visualized Entries



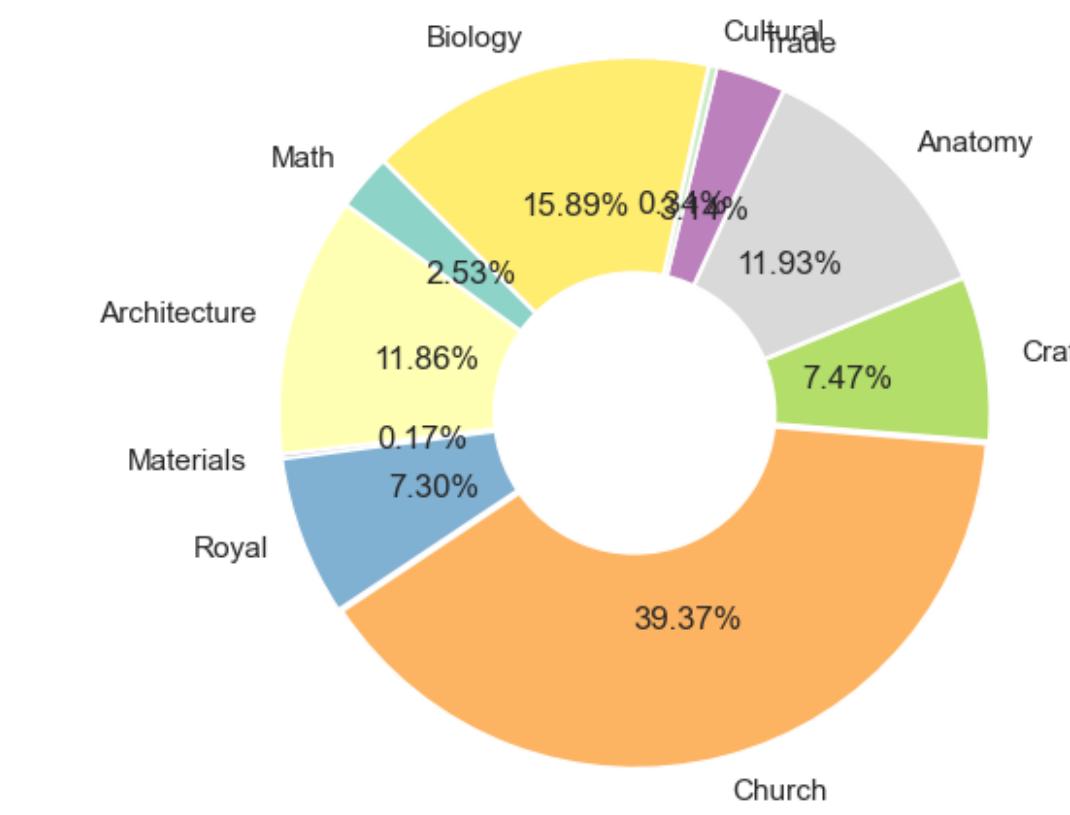
Topic Distribution in Version1



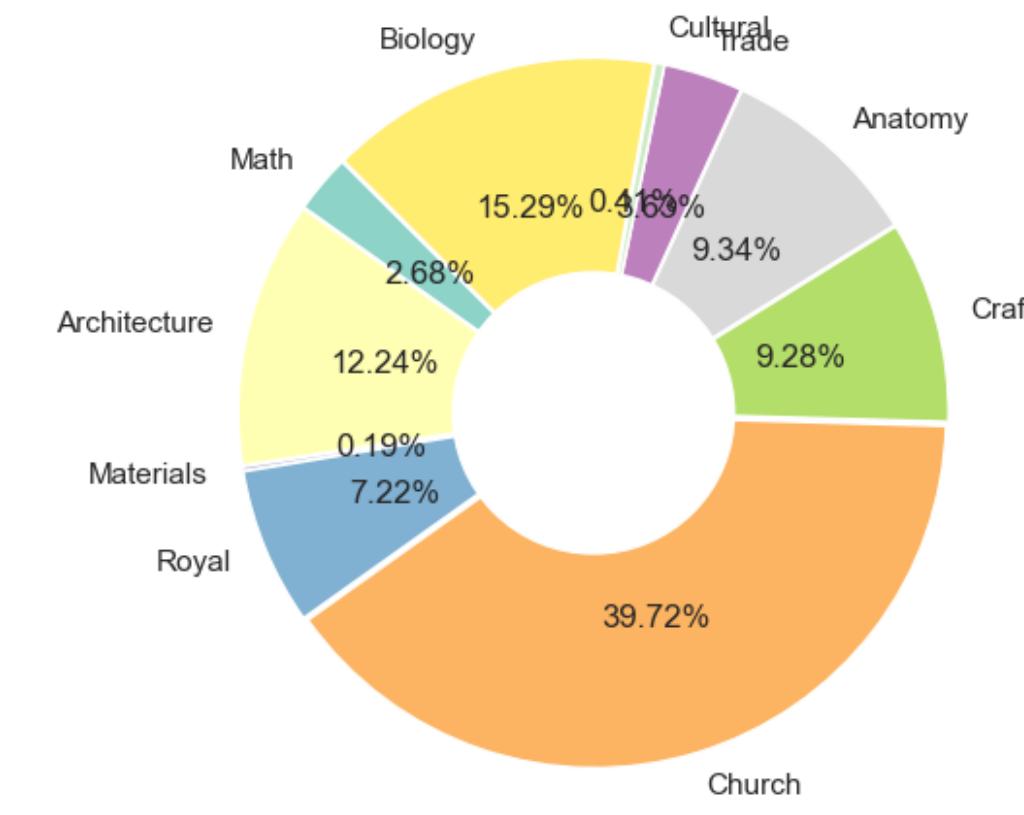
Topic Distribution in Version2



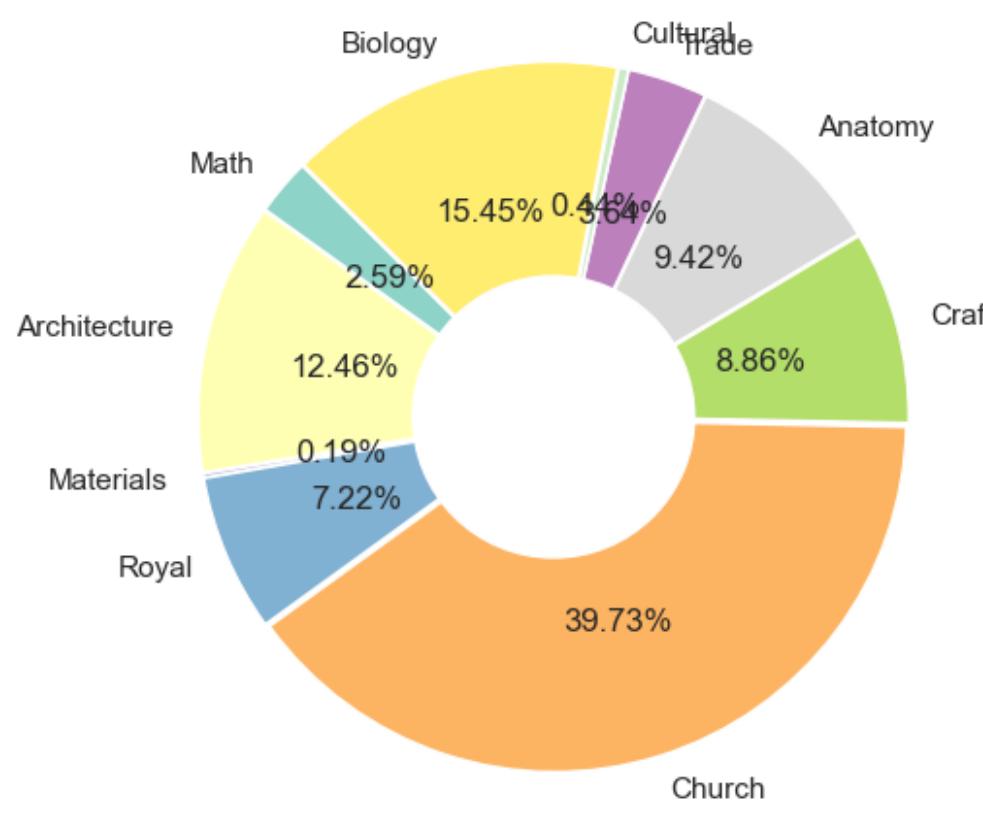
Topic Distribution in Version3



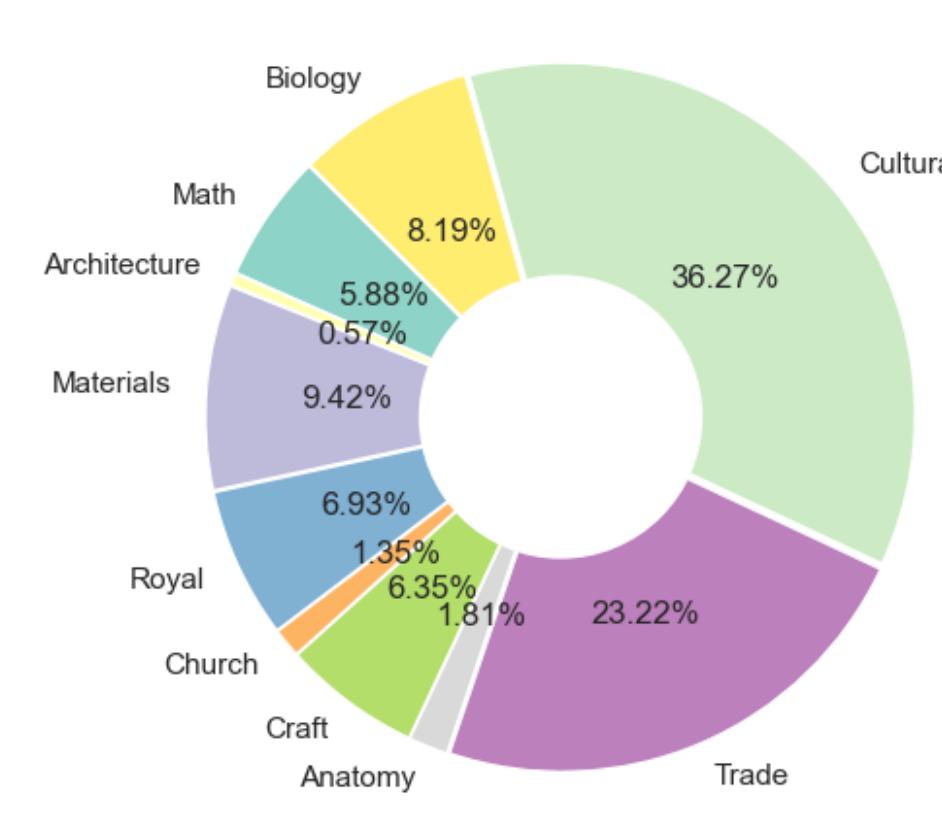
Topic Distribution in Version4



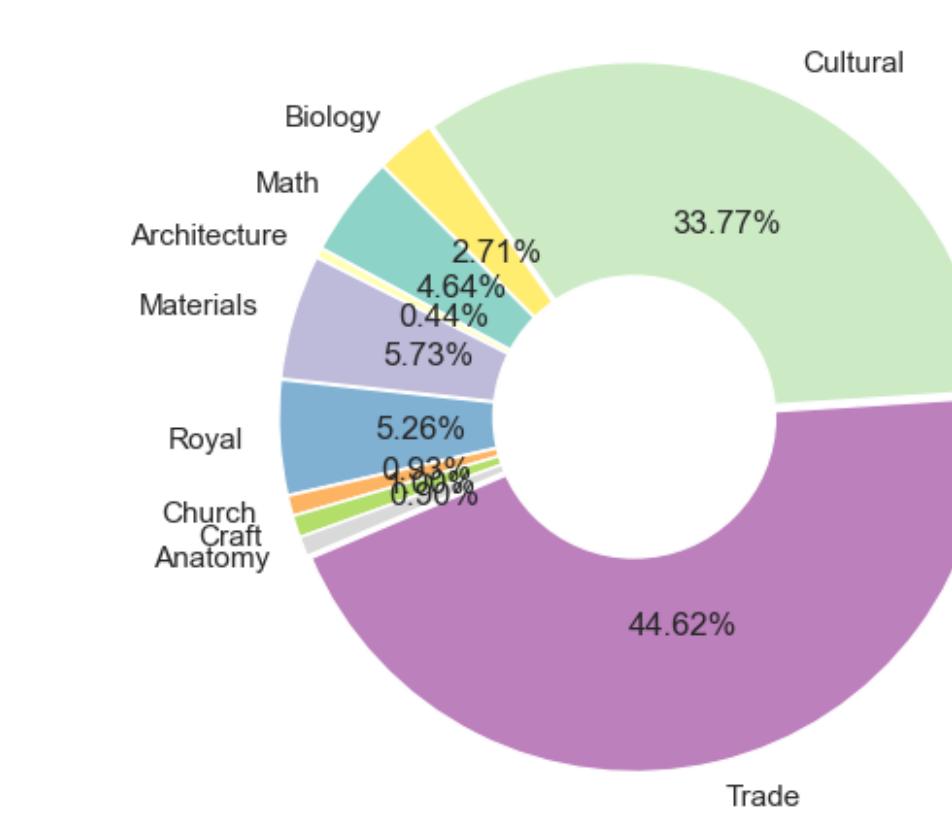
Topic Distribution in Version5



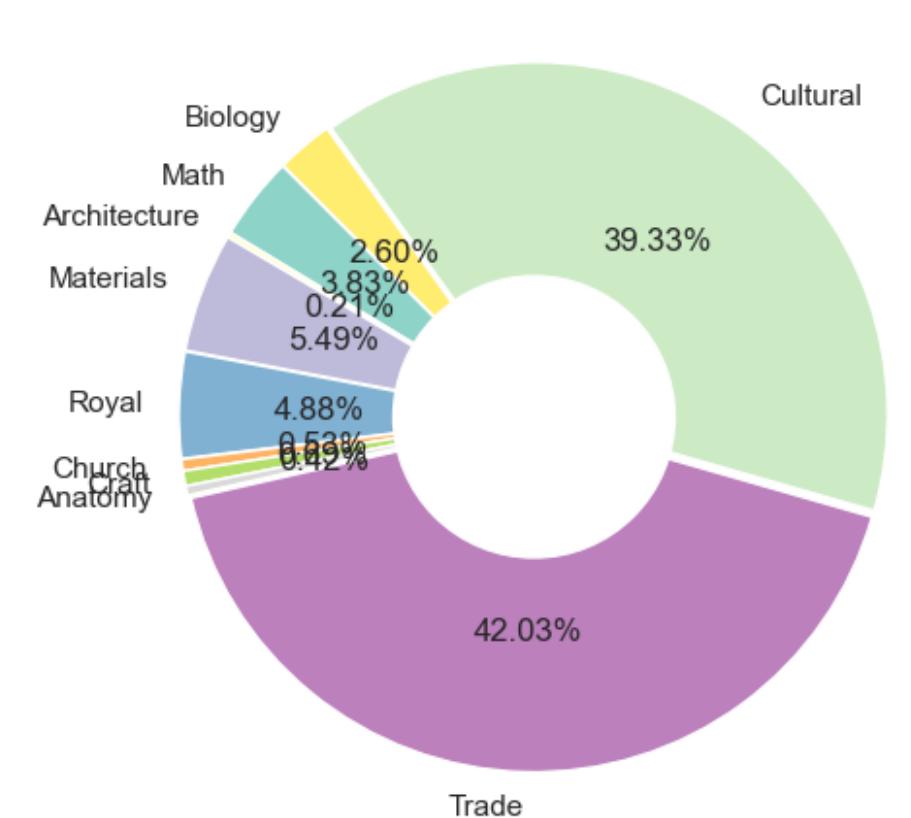
Topic Distribution in Version6



Topic Distribution in Version7



Topic Distribution in Version8



Percentage of Topics per Edition

Royal, Church, Craft, Trade, Cultural, Biology

Topic 3: Materials science

Topic 7: Anatomy

Topic 4: Military & Royal Family

Topic 8: Trade

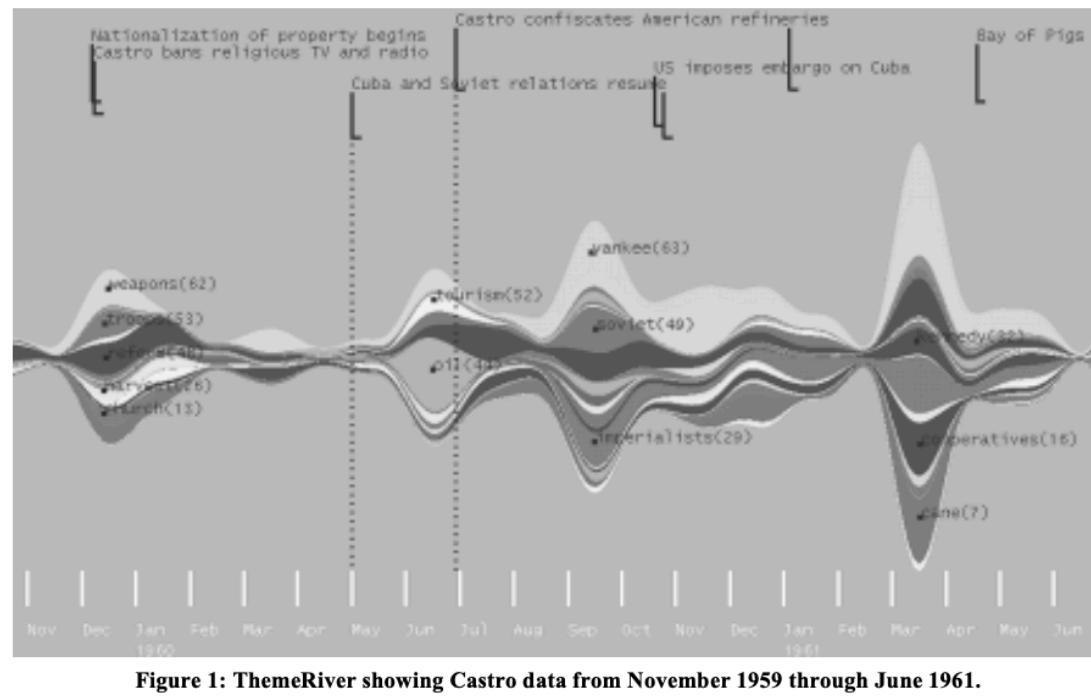
Topic 5: Politics & Church

Topic 9: Cultural

#5 Theme river

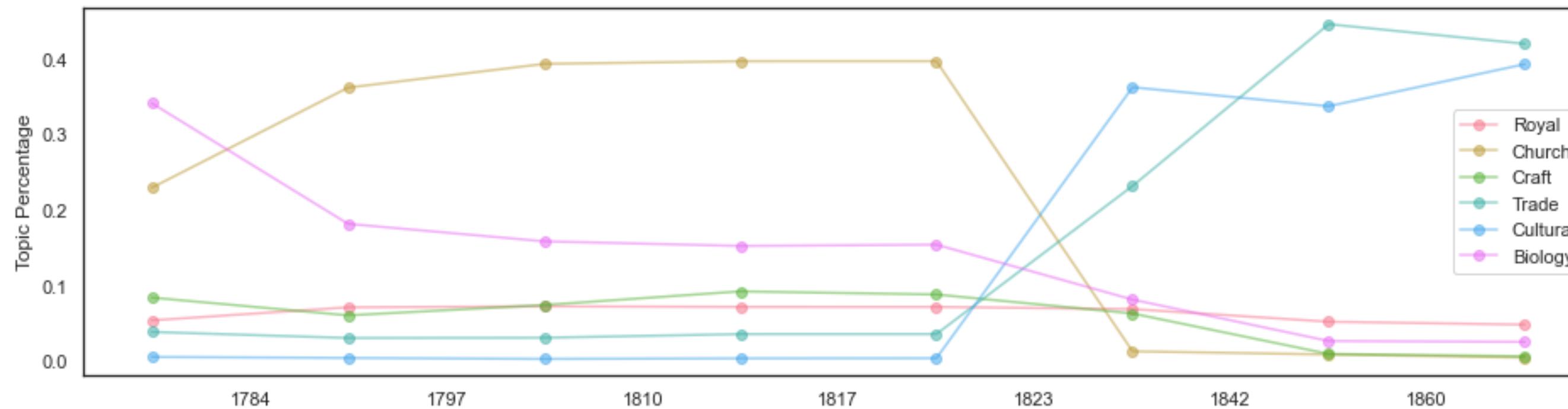
ThemeRiver™: In Search of Trends, Patterns, and Relationships

Susan Havre, Beth Hetzler, and Lucy Nowell
Battelle Pacific Northwest Division
Richland, Washington 99352 USA
1+509+375-6948



A. Reference

Havre, Susan L. et al. "ThemeRiver*: In Search of Trends, Patterns, and Relationships." (1999).



B. "Theme River" in Encyclopedia Britannica

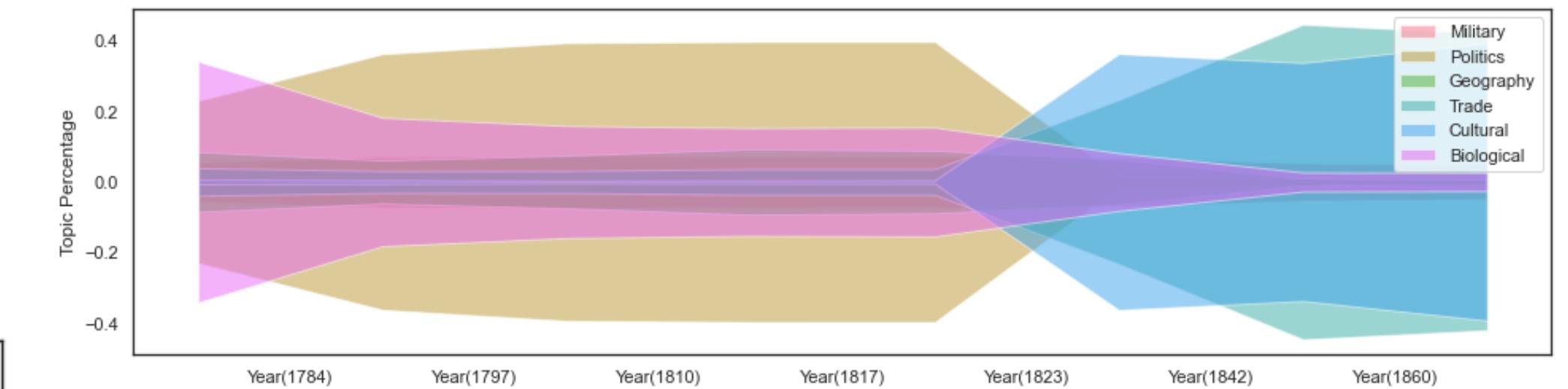
```
sns.set_theme(style="white", palette=sns.color_palette("husl"))

topic_labels=["Math", "Spots", "Chemical", "Military", "Politics", "Geography", "Agricultural", "Trade"]

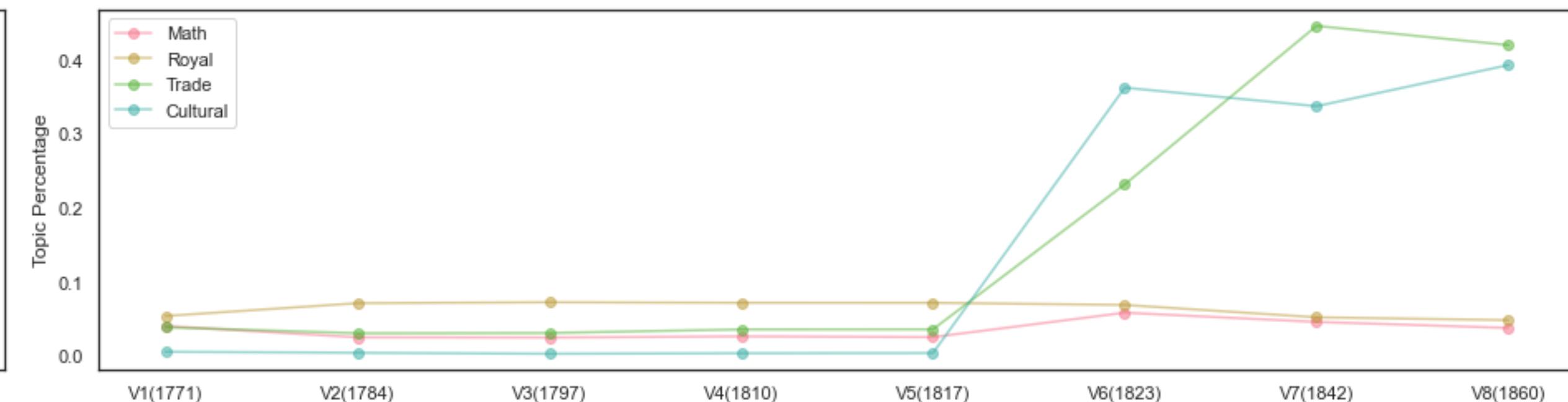
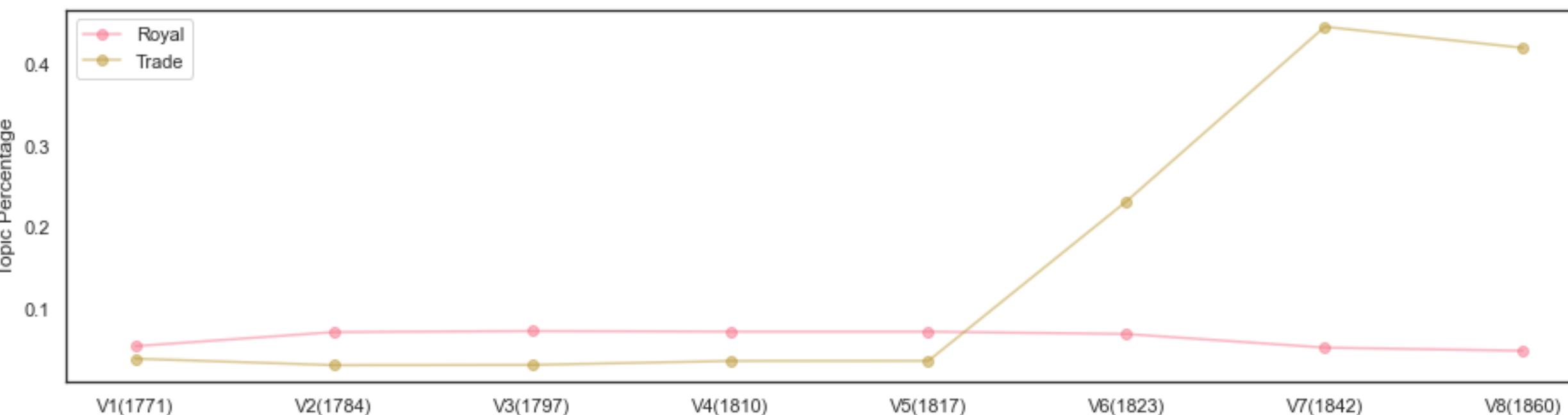
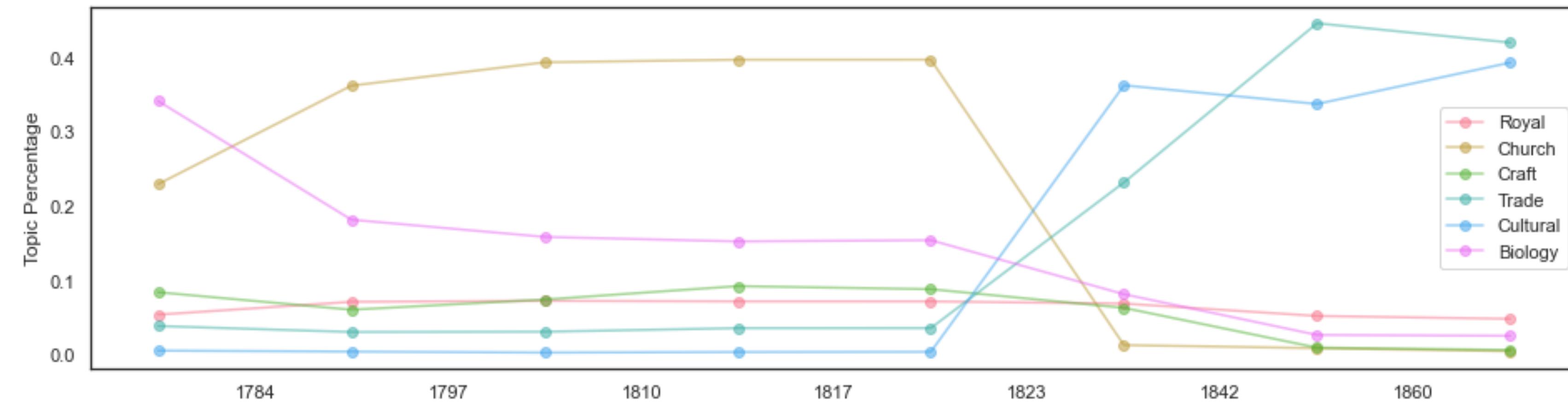
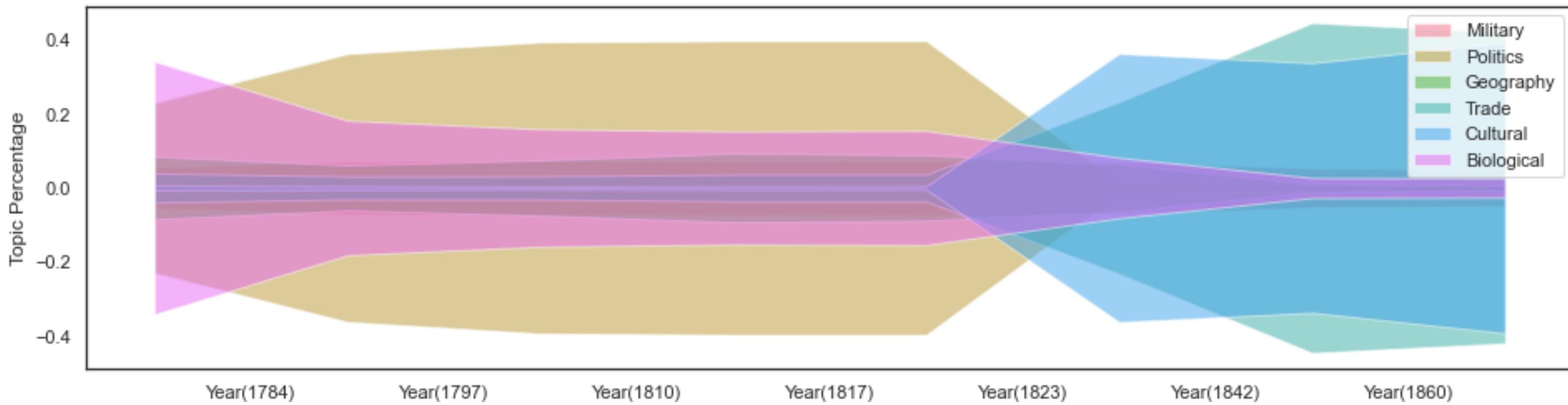
df_topic_pct=df_version_topic.div(df_version_topic.sum(axis=1), axis=0)
plt.figure(figsize=(16,4))
plt.xticks([x+0.5 for x in range(8)],
           ['Year(1771)', 'Year(1784)', 'Year(1797)', 'Year(1810)', 'Year(1817)', 'Year(1823)', 'Year(1842)', 'Year(1860)'],
           rotation=0)

for i in [3,4,5,7,8,9]: # range(10) 所有的话就range(10)
    plt.fill_between(x=df_topic_pct.index,
                     y1=df_topic_pct.iloc[:,i],
                     y2=-1*df_topic_pct.iloc[:,i],
                     alpha=0.5, label=topic_labels[i])

plt.legend()
plt.ylabel("Topic Percentage", fontsize=12)
```



C. Theme change line chart



#6 Interactive Board - Arduino and Pressure Sensor





Encyclopedie Britannica

— Topics in LDA model —



1775-1779

The Continuing War



1775 American Revolution



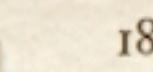
1778 Anglo-French War



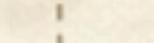
1779 The English-Spanish War



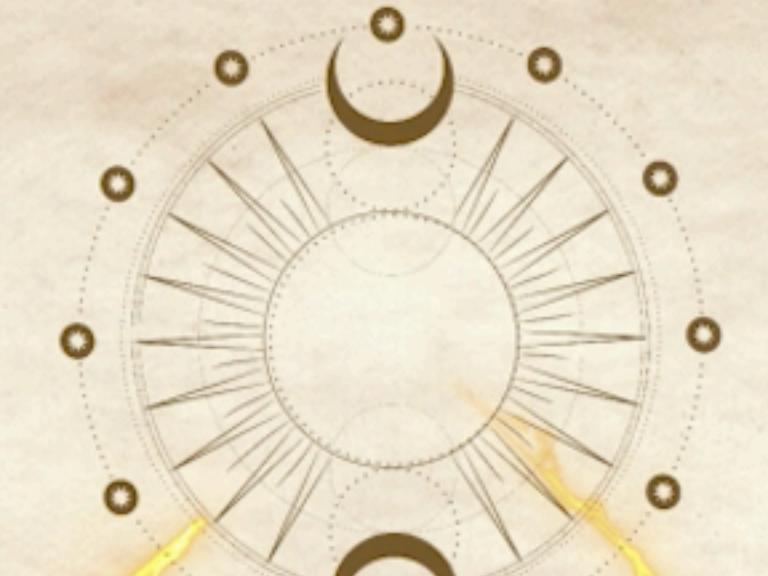
Soon after the imperial dominion of France was conferred on Napoleon Bonaparte, the Dutch republic was elevated to the rank of a kingdom, and the emperor's brother, Louis, was appointed the first king of Holland.



1832
Reform Decree
Giving political rights
to the industrial
bourgeoisie



1825
World's first public
passenger railway, the
Stockton &
Darlington Railway,
opens



1840
The First Opium War



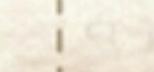
A commission
appointed by the
English government
in 1848 with the view
of promoting the
public health in
England and Wales,
which, in that and the
preceding year had
suffered seriously
from the invasion of
cholera.



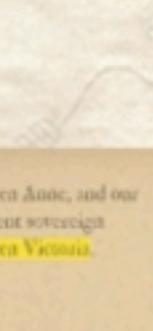
1810
From hand production
methods to machines



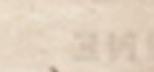
The Queen of
England is either
queen regent, queen
consort, or queen
dowager. The queen
regent, regnant, or
sovereign, is she who
holds the crown in
her own right, as
Queen Elizabeth.



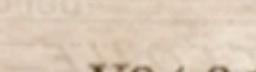
1845
The Great Irish Famine



Queen Anne, and our
present sovereign
Queen Victoria.



V7 (1845)



V8 (1860)

ROYAL
CHURCH
CRAFT
TRADE
CULTURAL
BIOLOGY

V2 (1784)

V3 (1797)

V4 (1810)

V5 (1817)

V6 (1823)

V7 (1845)

V8 (1860)

#7 Modeling Making

