DS4D Group 14

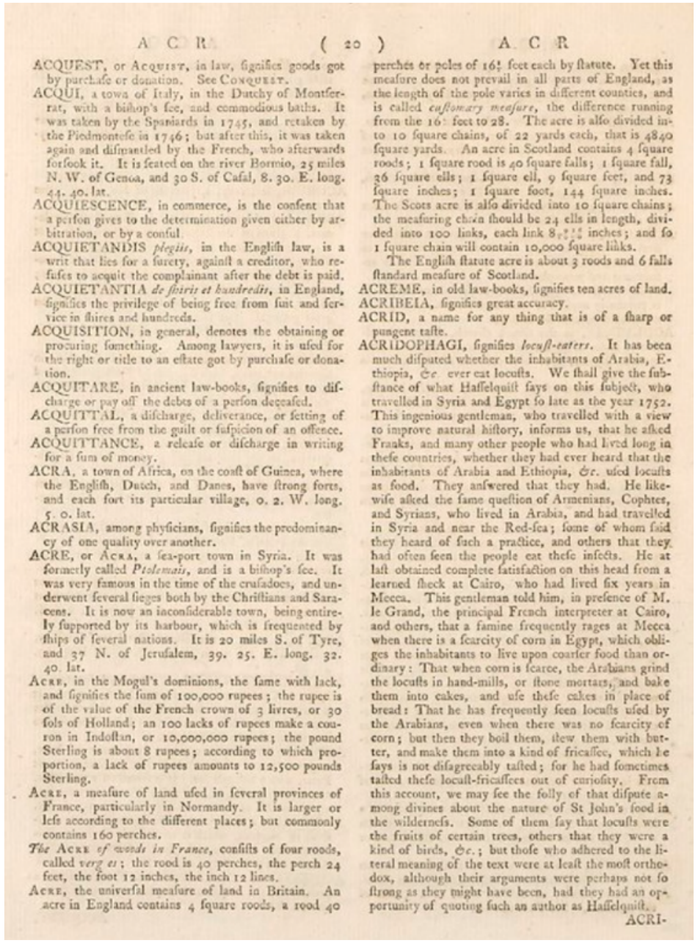
Encyclopedia Britannica

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



A C R (20) A C R

ACQUEST, or ACQUEST, in law, ſignifies goods got by purchaſe or donation. See CONQUEST.
ACQUI, a town of Italy, in the Dutchy of Monferrat, with a biſhop's ſee, and commodious baths. It was taken by the Spaniards in 1745, and retaken by the Piedmonteſe in 1746; but after this, it was taken again and diſmantled by the French, who afterwards forſook it. It is ſeated on the river Bormio, 25 miles N. W. of Genoa, and 30 S. of Caſal, 8. 30. E. long. 44. 40. lat.
ACQUIESCENCE, in commerce, is the conſent that a perſon gives to the determination given either by arbitration, or by a conſul.
ACQUIETANDIS *plegiis*, in the Engliſh law, is a writ that lies for a ſurety, againſt a creditor, who refuſes to acquit the complainant after the debt is paid.
ACQUIETANTIA *de ſhiris et hundredis*, in England, ſignifies the privilege of being free from ſuit and ſervice in ſhires and hundreds.
ACQUISITION, in general, denotes the obtaining or procuring ſomething. Among lawyers, it is uſed for the right or title to an eſtate got by purchaſe or donation.
ACQUITTARE, in ancient law-books, ſignifies to diſcharge or pay off the debts of a perſon deceaſed.
ACQUITTAL, a diſcharge, deliverance, or ſetting of a perſon free from the guilt or ſuſpicion of an offence.
ACQUITTANCE, a releaſe or diſcharge in writing for a ſum of money.
ACRA, a town of Africa, on the coaſt of Guinea, where the Engliſh, Dutch, and Danes, have ſtrong forts, and each fort its particular village, 0. 2. W. long. 5. 0. lat.
ACRASIA, among phyſicians, ſignifies the predominancy of one quality over another.
ACRE, or ACRA, a ſea-port town in Syria. It was formerly called *Ptolemais*, and is a biſhop's ſee. It was very famous in the time of the cruſadoes, and underwent ſeveral ſieges both by the Chriſtians and Saracens. It is now an inconſiderable town, being entirely ſupported by its harbour, which is frequented by ſhips of ſeveral nations. It is 20 miles S. of Tyre, and 37 N. of Jeruſalem, 39. 25. E. long. 32. 40. lat.
ACRE, in the Mogul's dominions, the ſame with lack, and ſignifies the ſum of 100,000 rupees; the rupee is of the value of the French crown of 3 livres, or 30 ſols of Holland; an 100 lacks of rupees make a coûron in Indoſtan, or 10,000,000 rupees; the pound Sterling is about 8 rupees; according to which proportion, a lack of rupees amounts to 12,500 pounds Sterling.
ACRE, a meaſure of land uſed in ſeveral provinces of France, particularly in Normandy. It is larger or leſs according to the different places; but commonly contains 160 perches.
The ACRE *of woods in France*, conſiſts of four roods, called *verges*; the rood is 40 perches, the perch 24 feet, the foot 12 inches, the inch 12 lines.
ACRE, the univerſal meaſure of land in Britain. An acre in England contains 4 ſquare roods, a rood 40 perches or poles of 16½ feet each by ſtatute. Yet this meaſure does not prevail in all parts of England, as the length of the pole varies in different counties, and is called *cuſtomary meaſure*, the difference running from the 16½ feet to 28. The acre is alſo divided into 10 ſquare chains, of 22 yards each, that is 4840 ſquare yards. An acre in Scotland contains 4 ſquare roods; 1 ſquare rood is 40 ſquare falls; 1 ſquare fall, 36 ſquare ells; 1 ſquare ell, 9 ſquare feet, and 73 ſquare inches; 1 ſquare foot, 144 ſquare inches. The Scots acre is alſo divided into 10 ſquare chains; the meaſuring chain ſhould be 24 ells in length, divided into 100 links, each link $8\frac{37}{100}$ inches; and ſo 1 ſquare chain will contain 10,000 ſquare links.
The Engliſh ſtatute acre is about 3 roods and 6 falls ſtandard meaſure of Scotland.
ACREME, in old law-books, ſignifies ten acres of land.
ACRIBEIA, ſignifies great accuracy.
ACRID, a name for any thing that is of a ſharp or pungent taſte.
ACRIDOPHAGI, ſignifies *locuſt-eaters*. It has been much diſputed whether the inhabitants of Arabia, Ethiopia, &c. ever eat locuſts. We ſhall give the ſubſtance of what Haſſelquiſt ſays on this ſubject, who travelled in Syria and Egypt ſo late as the year 1752. This ingenious gentleman, who travelled with a view to improve natural hiſtory, informs us, that he aſked Franks, and many other people who had lived long in theſe countries, whether they had ever heard that the inhabitants of Arabia and Ethiopia, &c. uſed locuſts as food. They anſwered that they had. He likewiſe aſked the ſame queſtion of Armenians, Cophtes, and Syrians, who lived in Arabia, and had travelled in Syria and near the Red-ſea; ſome of whom ſaid they heard of ſuch a practice, and others that they had often ſeen the people eat theſe inſects. He at laſt obtained complete ſatisfaction on this head from a learned ſheck at Cairo, who had lived ſix years in Mecca. This gentleman told him, in preſence of M. le Grand, the principal French interpreter at Cairo, and others, that a famine frequently rages at Mecca when there is a ſcarcity of corn in Egypt, which obliges the inhabitants to live upon coarſer food than ordinary: That when corn is ſcarce, the Arabians grind the locuſts in hand-mills, or ſtone mortars, and bake them into cakes, and uſe theſe cakes in place of bread: That he has frequently ſeen locuſts uſed by the Arabians, even when there was no ſcarcity of corn; but then they boil them, ſtew them with butter, and make them into a kind of fricaſſee, which he ſays is not diſagreeably taſted; for he had ſometimes taſted theſe locuſt-fricaſſees out of curioſity. From this account, we may ſee the folly of that diſpute among divines about the nature of St John's food in the wilderneſs. Some of them ſay that locuſts were the fruits of certain trees, others that they were a kind of birds, &c.; but thoſe who adhered to the literal meaning of the text were at leaſt the moſt orthodox, although their arguments were perhaps not ſo ſtrong as they might have been, had they had an opportunity of quoting ſuch an author as Haſſelquiſt.

ACRI-

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

Text cleanup will restore the encyclopedia to its original structure

1. 清理出每个版本的词条

```
# 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.清理出词条

```
# 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', 'AB
RY', 'ABDALS', 'ABDELAVI', 'ABDEST', '
LIANS', 'ABELMOSCH', 'ABENSBURG', 'ABE
'ABERRATION', 'ABERYSWITH', 'ABESTA',
L', 'ABIB', 'ABIES', 'ABILITY', 'ABING
BLACTATION', 'ABLACQUEATION', 'ABLATIV
ION', 'ABO', 'ABOARD', 'ABOLITION', 'A
RTION', 'ABORTIVE', 'ABOY', 'ABRA', 'A
BRAUM', 'ABRASAX', 'ABRAX', 'ABREAST',
```

B.全部的词条

```
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 fpell, 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 againd agues 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 Ikin is fo tender and lax as to render them fubjeft to abrafion.	1
ABRASION	in medicine, the corroding 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
ABRASAX	or Abraxas, a myftical term found in the ancient theology and philofophy of Balilides's followers	1

C.汇总每个版本的词条

2. 进一步清理，去除OCR的杂音

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


def clean_text(text):

    # Remove all non-English letters directly
    patten=r'[^a-z]'
    text=re.sub(patten,' ',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.进一步清理

```
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 fpell, 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 againd agues andfeveral other difeaies ABRAHAM'r balm, in botany, See Cannabis.	1	magical word fpell written many times word contains letters omitting letter former every time ages ignorance fuperdition vvyrn neck antidote againd agues 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	1	order monks exterminated idolatry theophilus ninth century alfo name another fedt heretics

E.导出样例

Use LDA model

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

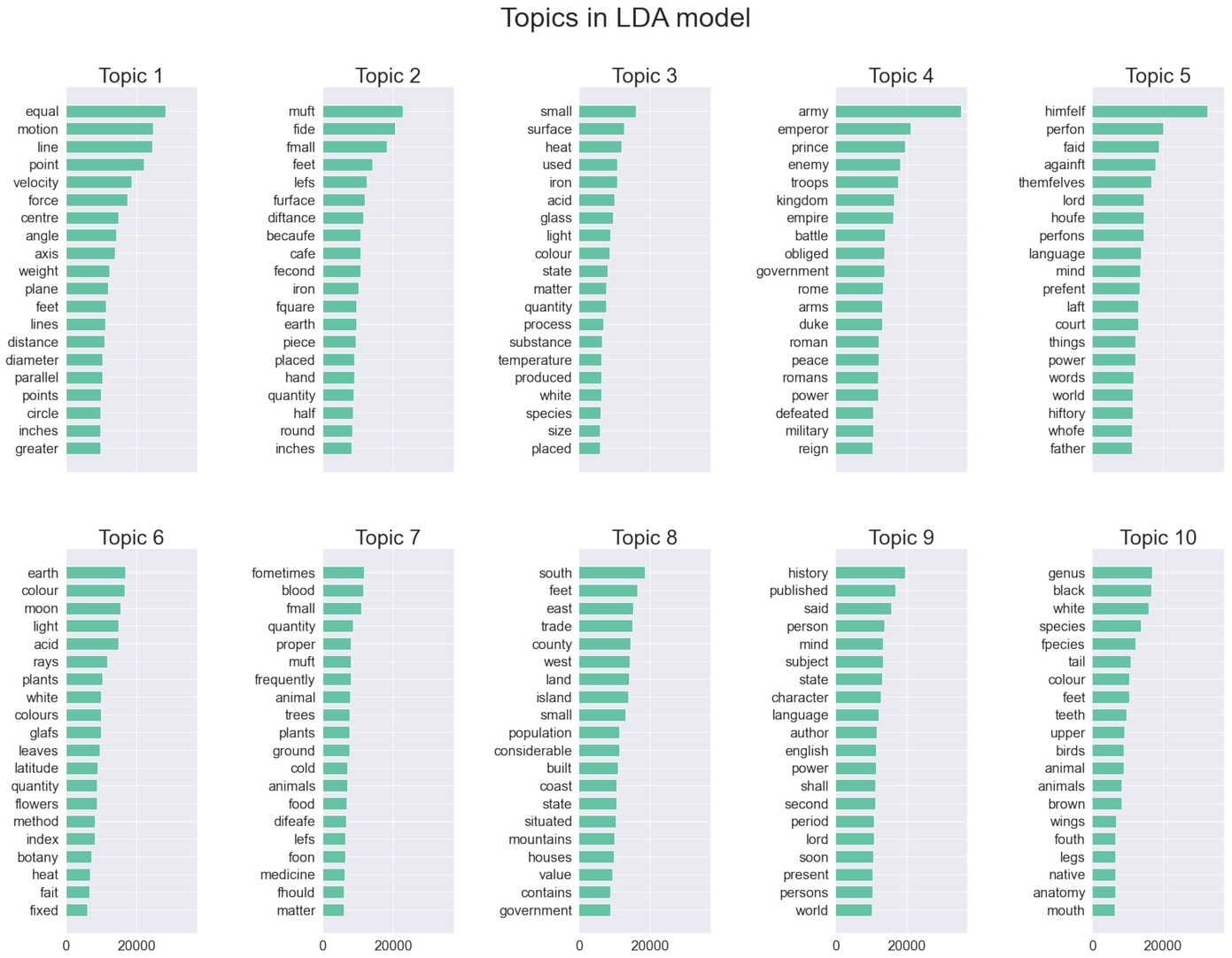
```
n_top_words = 20
feature_names = count.get_feature_names()

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:-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:
himfelf perfon faid againft themfelves lord houfe perfons language mind prefent laft court th
```

B.20个主题中的重要的词汇

#1 Using Topic Modeling and LDA to analyze Encyclopedia

```
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个主题和他们重要的词汇

#2 Name the topic with the example sentences extracted from the general vocabulary

```
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:
himfelf perfon faid againft themfelves lord houfe perfons language mind prefent laft court th
ings power words world hiftory 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 shal
l second period lord soon present persons world
Topic 10:
genus black white species fpecies tail colour feet teeth upper birds animal animals brown win
gs 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 para

Topic2:

        sample#1:  the art of carting and working lead, and ufing it in building.As this met
 al melts foon and with little heat, it is eafy to cart; it into figures of any kind, by runni

Topic3:

        sample#1:  in building, a composition of white marble pulverised, and mixed with pla
ster of lime y and the whole being sifted and wrought np with water, is to be used like commo
n plaster: this is called by Pliny marmoratum opusr 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 Prafut
aqus king of the Iceni in Britain, who by his will left the emperor and his own daughters co-
Topic5:

        sample#1:  (in Greekthe book'), a name applied by Chriftians, by way of eminence or
diftindion, to the | collodion of facred writings, or the holy fcriptures of | the Old and Ne
w Teftaments; known alfo by various ' other appellations, as, the Sacred Books, Holy Writ, ,1
Infpired Writings, Scriptures, <bc. The Jews ftyled i the Bible (that is, the Old Teftament)
mikra; which fignifies Lcjfon, or Lefture.This collodion of the facred writings, containing |
thofe of the Old and New Teftament, is juftly looked ...

        sample#2:  a name applied by Chriftians, by way of eminence or diftindtion, to the c
olleftion of facred writings, or the holy fcriptures of the Old and New Teftament; known alfo
by various other appellations, as, the Sacred Books, Holy Writ, Infpired Writings, Scripture
s, 'be. The Jews ftiled the Bible (that is, the Old Telkment) fnikra, which fignifies Lejfon,
or Leflure. , This collection of the facred writings, containing thofe of the Old and New Tef
tament, is judly looked upon as the foundation of the ...

        sample#3:  a general name given to the body of ecclefiaftics of the Chriftian churc
h, in contradiftimftion to the laity. See Laity.The diftindtion of Chriftians into clergy and
```

#2 Name each topic

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 uas 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 ...

#3 Count the distribution of entries for each topic by version

```
# 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)
```

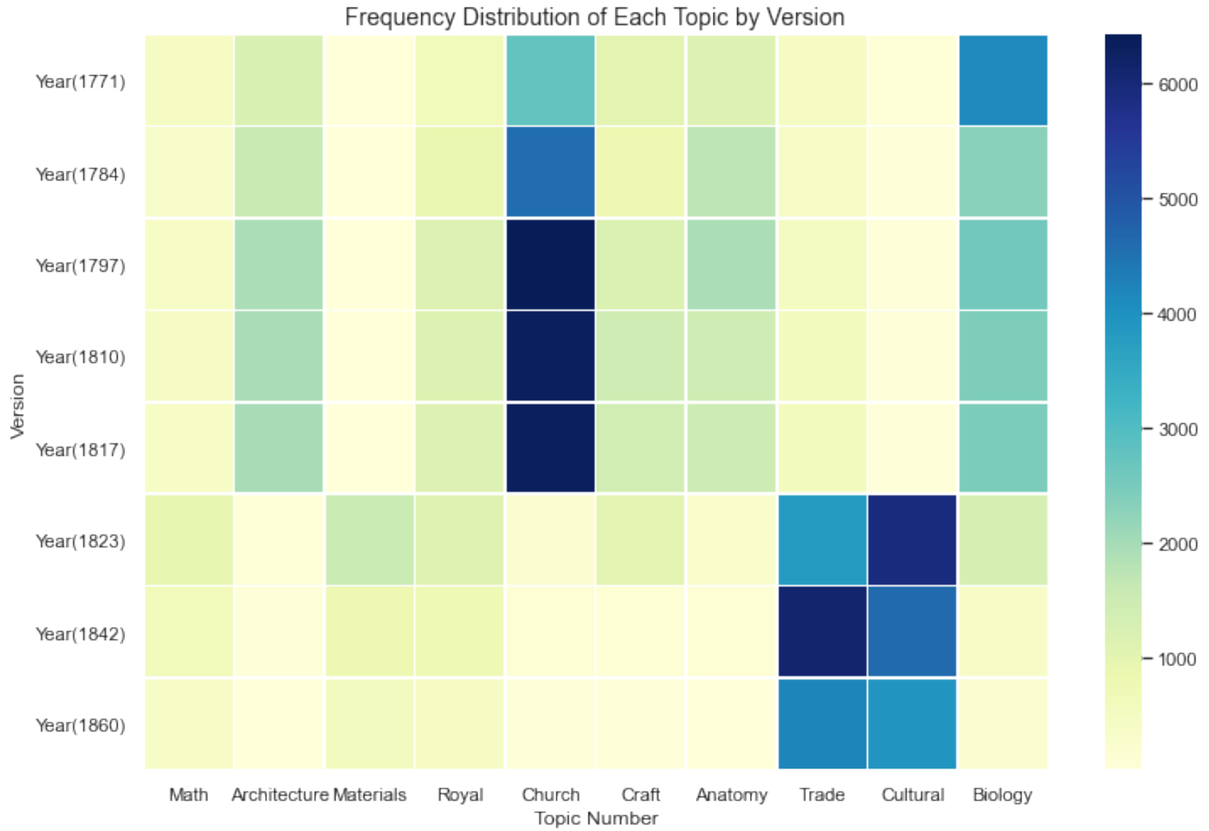
topic	0	1	2	3	4	5	6	7	8	9
version										
1	0.040850	0.105102	0.002563	0.054412	0.229968	0.084512	0.095923	0.039114	0.006037	0.341520
2	0.025595	0.123563	0.002205	0.071665	0.362262	0.061033	0.135769	0.031107	0.004725	0.182076
3	0.025297	0.118584	0.001715	0.072951	0.393667	0.074727	0.119258	0.031422	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(18
           rotation=0)

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

```
Text(0.5, 1.0, 'Frequency Distribution of Each Topic by Version')
```

A. Percentage of Visualized Entries



#4 “Theme river” & Reference

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

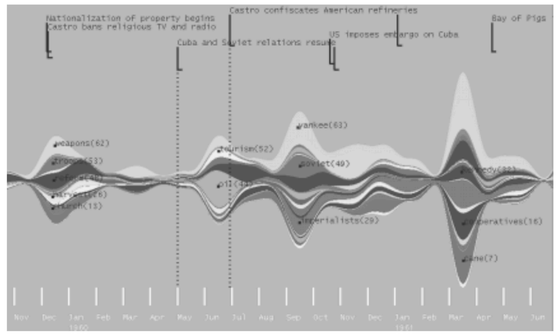


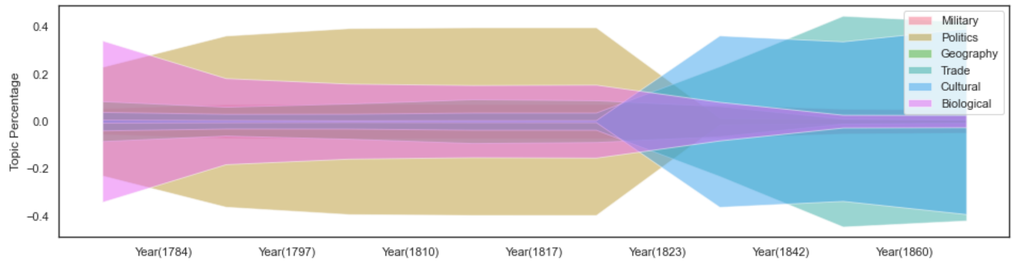
Figure 1: ThemeRiver showing Castro data from November 1959 through June 1961.

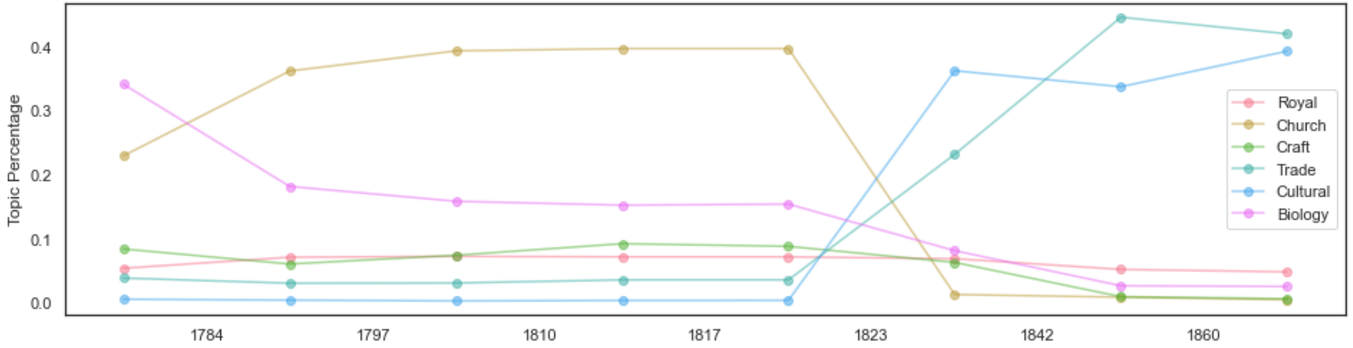
A. Reference

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

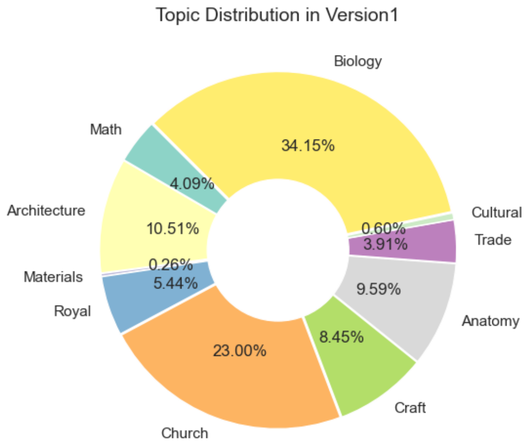
B. "Theme River" in Encyclopedia Britannica

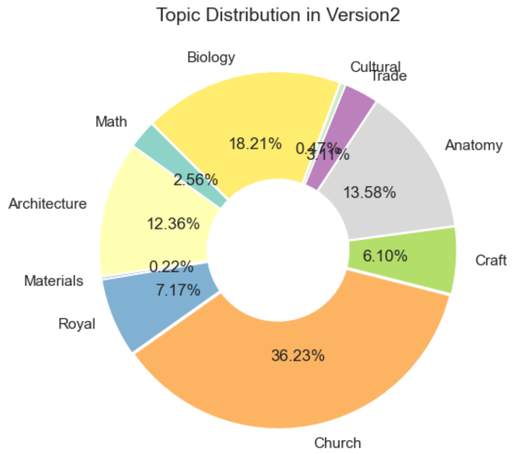
```
sns.set_theme(style="white",palette=sns.color_palette("husl"))

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
           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)
```

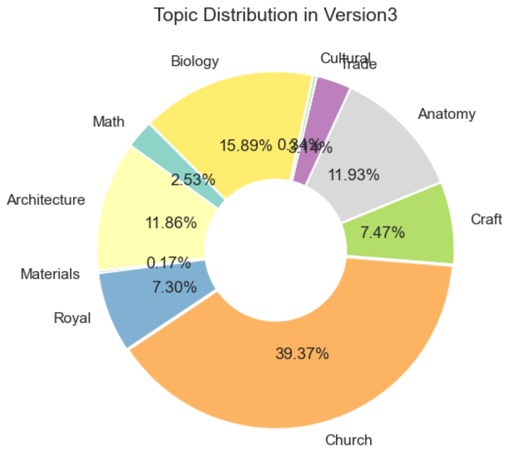


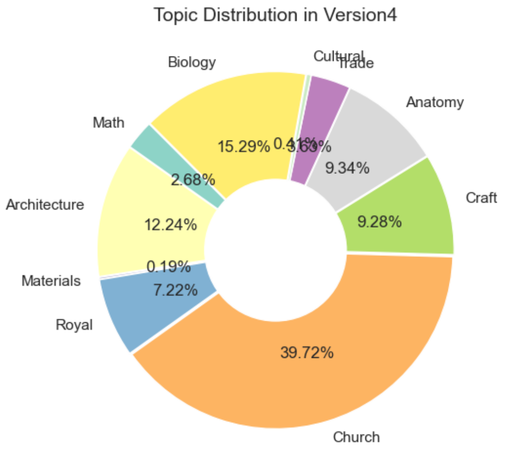


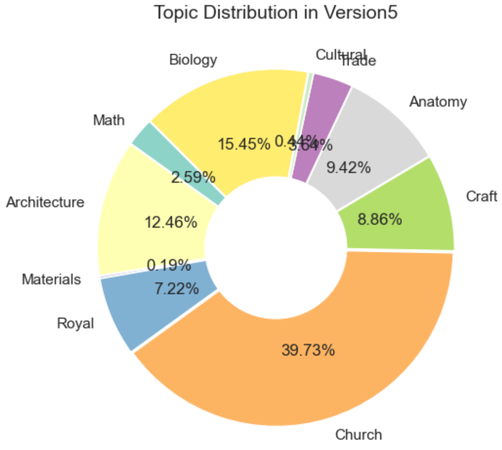
C. Theme change line chart

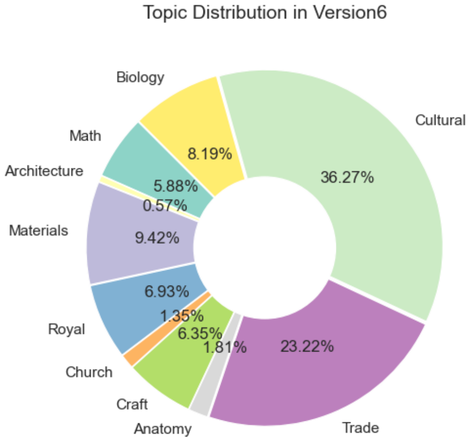


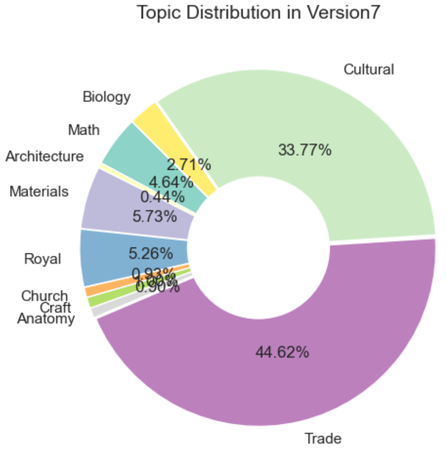


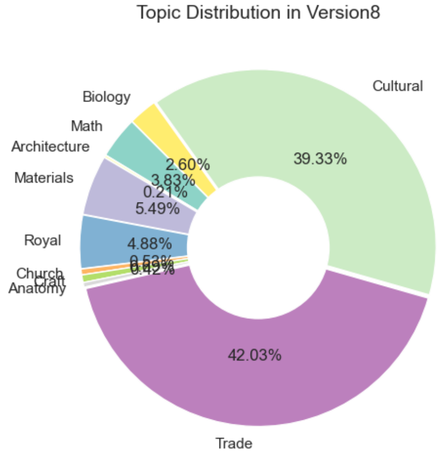






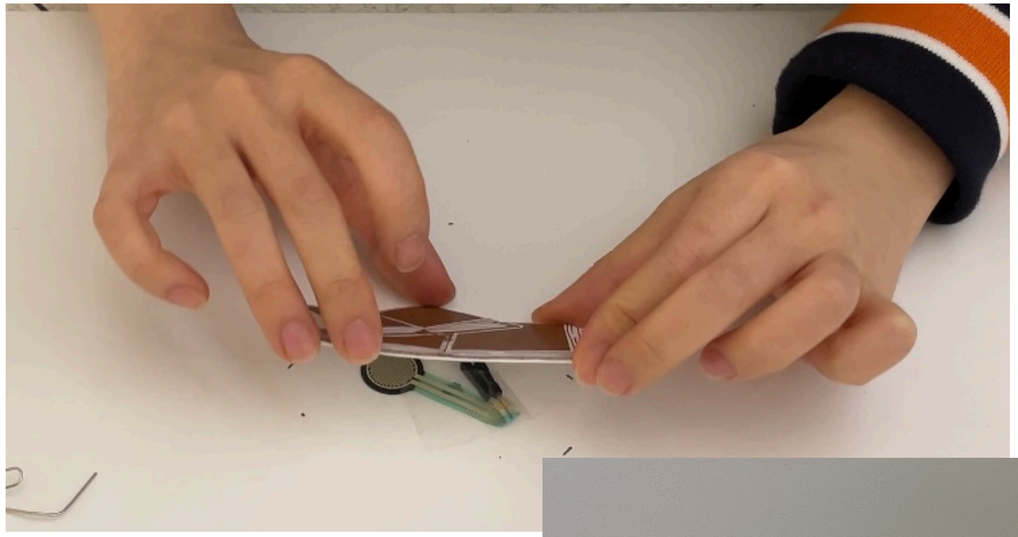


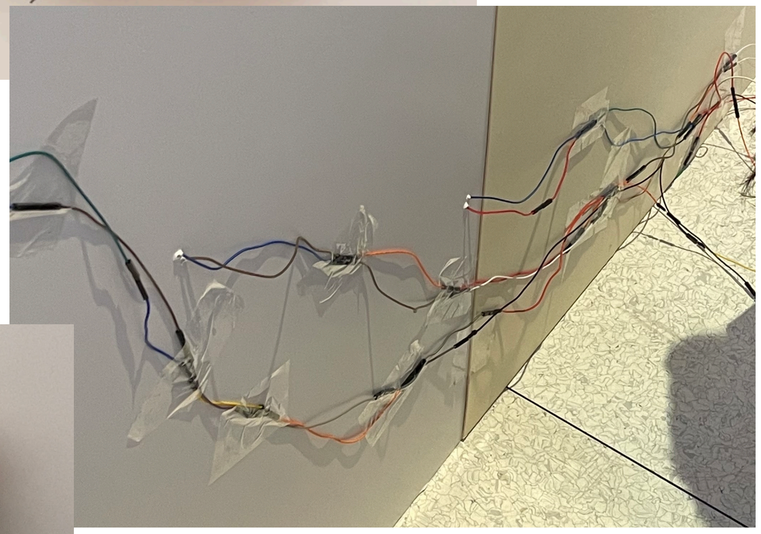


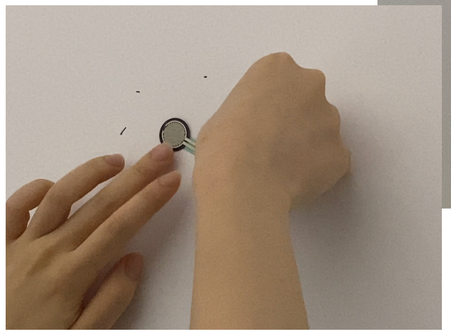


Percentage of Topics per Edition

#4 制作过程-压力传感器





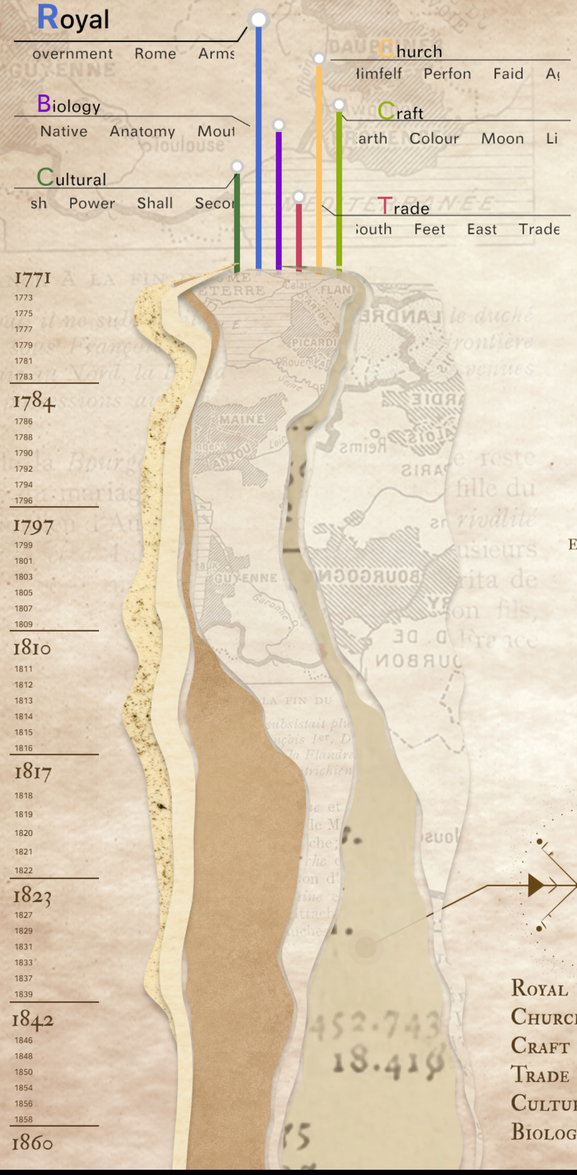


发现了?

- 补一张场景图

Encyclopedia Britannica

— Topics in LDA model —



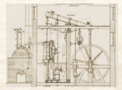
1775-1779

The Continuing War

- 1775 American War of Independence
- 1778 Anglo-French War
- 1779 The English-Spanish War

1772

East India Company in economic crisis



1769

Watt improves the steam engine

1815

Britain defeats Napoleonic dynasty and becomes the Empire of the Sun

1810

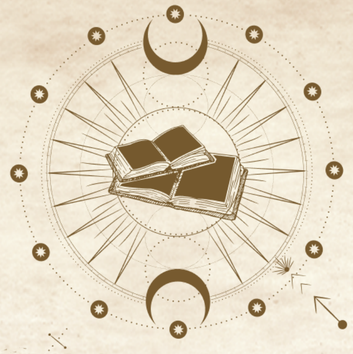
World's first public passenger railway, the Stoughton & Darlington Railway, opens

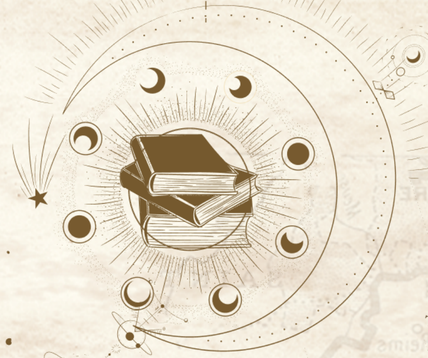
1832

Reform Decree
Giving political rights to the industrial bourgeoisie

1825

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





1840

The First Opium War

1845

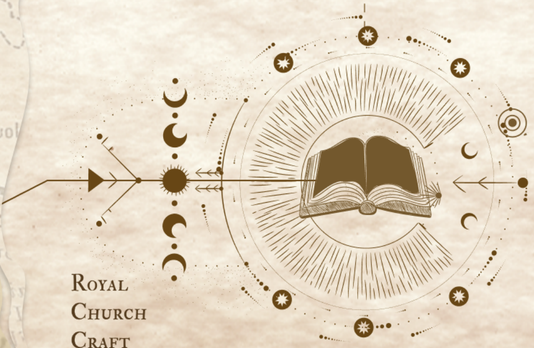
The Great Irish Famine

1848

Cholera epidemic

1837

Queen Victoria's succession to the throne
Significant population growth and increased labour force led to economic prosperity



Royal
Church
Craft
Trade
Cultural
Biology

1771 1784 1797 1810 1817 1823 1845 1860