

# Encyclopaedia Britannica

DS4D Group 14

Visual Interactive Projection Installation

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

ACQUEST, or Acquisit, in law, signifies goods got by purchase or donation. See CONQUEST.

ACQUI, a town of Italy, in the Duchy of Montferrat, 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.

ACQUITANDIS *plicatio*, 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 fiscis et hundredis*, in England, signifies the privilege of being free from suit and service in fiscs 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.

ACQUITARE, 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.

ACQUITTANCIS, 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 crusades, 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 couron 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 words in France consists of four rods, called *verg e*; the rod is 40 perches, the perch 24 feet, the foot 12 inches, the inch 12 lines.

# 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 ACQUISIT., is law, signifies goods got by purchase or donation. See CONQUEST.  
ACQUIU, a town of Italy, in the Duchy of Monferrat, with a bishop's see, and commodious baths. It was taken by the Spaniards in 1745, and retaken by the Piedmontese in 1761; but after this, it was taken again and dismantled by the French, who afterwards took it. It is seated on the river Borro, 25 miles N. W. of Genoa, and 30 S. of Cafal, 8. 30. E. long.  
ACQUIESCENCE, in commerce, is the consent that a person gives to the determination given either by arbitration, or by a consul.  
ACQUIETANDIS *plegis*, 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 ffrir et bandredit*, England, signifies the privilege of being free from suit and service in shires and hundred.  
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.  
ACQUITARE, in ancient law-books, signifies to discharge or pay off the debts of a person defeated.  
ACQUITTAL, a discharge, deliverance, or letting 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, o. 2. W. long.  
ACRASIA, among physicians, signifies the predominancy of one quality over another.  
ACRE, or ACRA, a seaport town in Syria. It was formerly called *Pisidemus*, and is a bishop's see. It was very famous in the time of the crusades, 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 Mayor's domains, the same with lack, and signifies the sum of 1200 pounds sterling; the sum of the value of the French crown of 3 litres, or 30 sols of Holland; an 100 lacks of rupees make a canon in Indofan, 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 meaure 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, or acre, in France, consists of four roads, called *voye*; the road is 40 perches, the perch 24 feet, the foot 12 inches, and the inch 12 lines.*  
Acre, the universal meaure of land in Britain. An acre in England contains 4 square rods, a rod 40

## 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
```

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

dict_keys(['SCO', 'AABAM', 'AADE',
ACATUATA', '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', 'AB
LIANS', 'ABELMOSCH', 'ABENSBURG', 'ABE
'ABERRATION', 'ABERYSWITH', 'ABESTA',
L', 'ABIB', 'ABIES', 'ABILITY', 'ABING
BLACTATION', 'ABLACQUEATION', 'ABLATIVE
ION', 'ABO', 'ABOARD', 'ABOLITION', 'AB
RTION', 'ABORTIVE', 'ABOY', 'ABRA', 'AB
RAUM', 'ABRASAX', 'ABRAX', 'ABREAST'.
```

df_encyclopedia_all.iloc[100:110,:]		
	text	version
ABRA	a silver coin of Poland, in value nearly equivalent to an Engliifi Hulling.	1
ABRACADABRA	a magical word or spell, which being written as many times as the word contains letters, and omitting the last letter of the former every time, was, in the ages of ignorance and superstition, worn about the neck, as an antidote against agues and several other diseases ABRAHAM's balm, in botany. See Cannabis.	1
ABRAHAMITES	an order of monks exterminated - for idolatry by Theophilus in the ninth century. Also the name of another sect of heretics who had adopted the errors of Paulus. See Paulicians.	1
ABRAMIS	an obsolete name for the fish cyprinus. See Cyprinus, B Abrasa, A B R (6 ABRASA, in surgery, ulcers, where the skin is so tender and lax as to render them subject to abrasion.	1
ABRASION	in medicine, the corroding of any part by acrid humours or medicines.	1
ABRAUM	an obsolete name of a certain species of clay, called by some authors Adamite earth, on account of its red colour.	1
ABRASAY	or Abraxas, a mystical term found in the ancient theology and philosophy of Pythagoras's followers.	1

A. Clean up entries

B. All entries

C. Collect all entries of each edition

## 2. Clean OCR's noise

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

def clean_text(text):

    # Remove all non-English letters directly
    pattern=r'[^\w\s]'

    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)
```

```
df_encyclopedia_all["text_clean"] = df_encyclopedia_all["text"].apply(clean_text)

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

	text	version	text_clean
ABRA	a silver coin of Poland, in value nearly equivalent to an Engliifi Hulling.	1	silver coin poland value nearly equivalent engliifi hulling
ABRACADABRA	a magical word or spell, which being written as many times as the word contains letters, and omitting the last letter of the former every time, was, in the ages of ignorance and superstition, worn about the neck, as an antidote against agues and several other diseases ABRAHAM's balm, in botany. See Cannabis.	1	magical word spell written many times word contains letters omitting letter former every time ages ignorance superstition worn neck antidote against agues and several other diseases abraham balm botany cannabis
ABRAHAMITES	an order of monks exterminated - for idolatry by Theophilus in the ninth century. Also the name of another sect of heretics who had adopted the errors of Paulus. See Paulicians.	1	order monks exterminated idolatry theophilus ninth century also name another sect heretics

D. Keep cleaning

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:-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:  
muit fide fmaille feet lefs furface difntance 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 Important Vocabularies for Each Topic

## 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.bars(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 3 Materials science
- Topic 4 Military & Royal Family
- Topic 5 Politics & Church
- Topic 6 Anatomy
- Topic 7 Trade
- Topic 8 Cultural

## 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**

1.

### 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<sup>^</sup> 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 ...

2.

### 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. 430. north latitude, and ...

3.

### 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

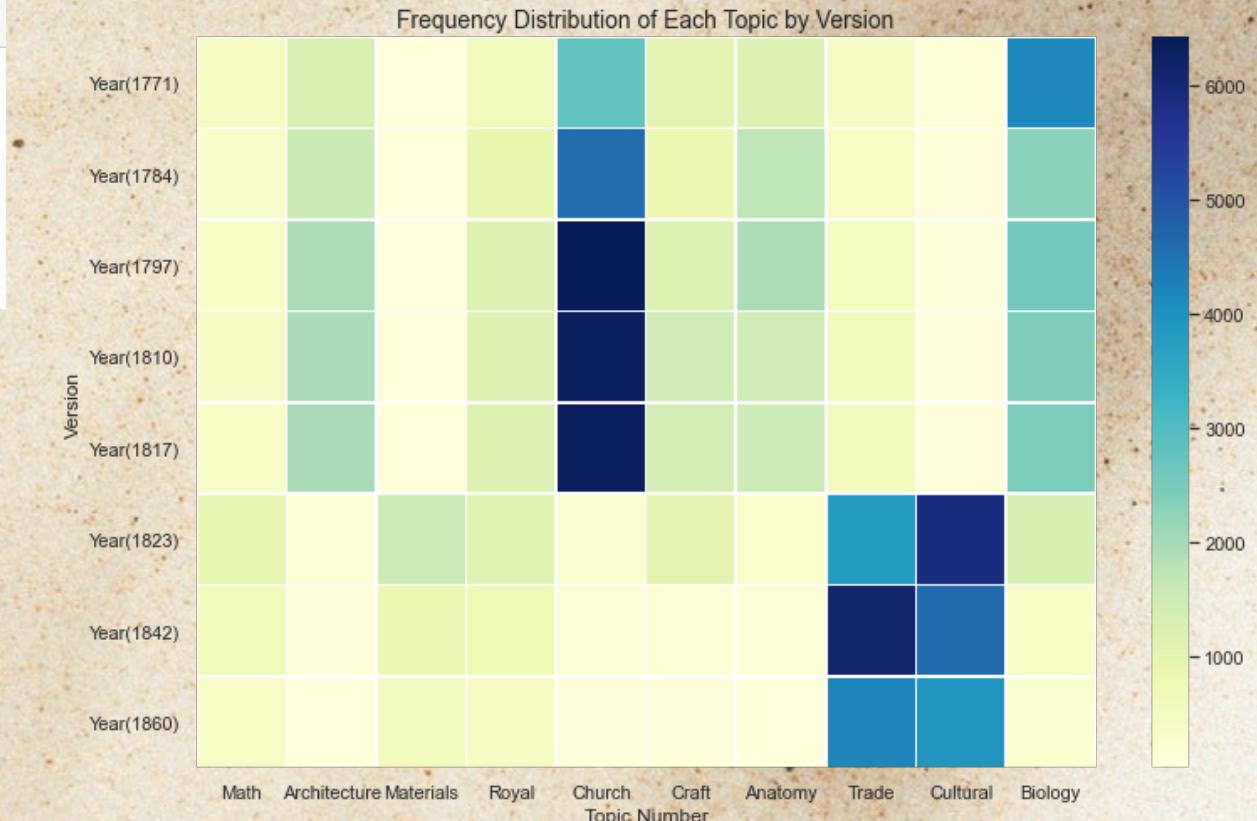
```
# Name the theme and visualize it
topic_labels=["Math", "Architecture", "Materials", "Royal", "Church", "Craft", "Anatomy", "T
plt.figure(figsize=(12,8))
sns.heatmap(df_version_topic, cmap="YlGnBu", linewidths=.5)

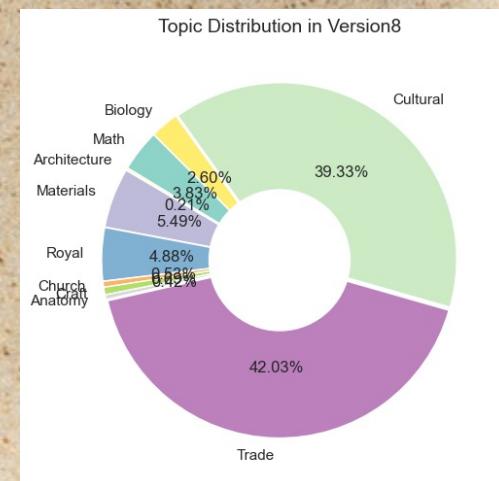
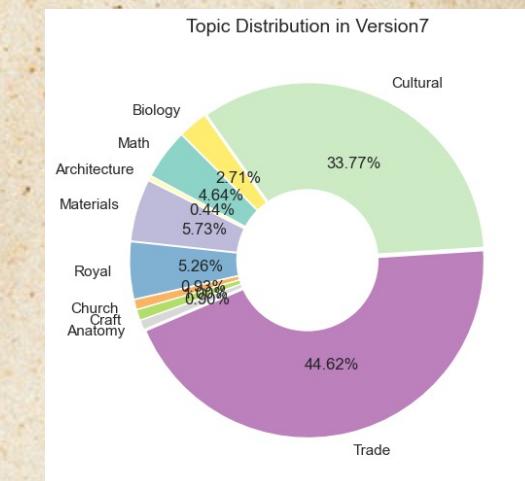
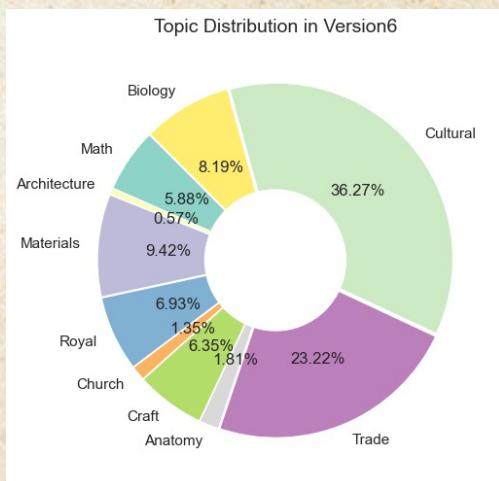
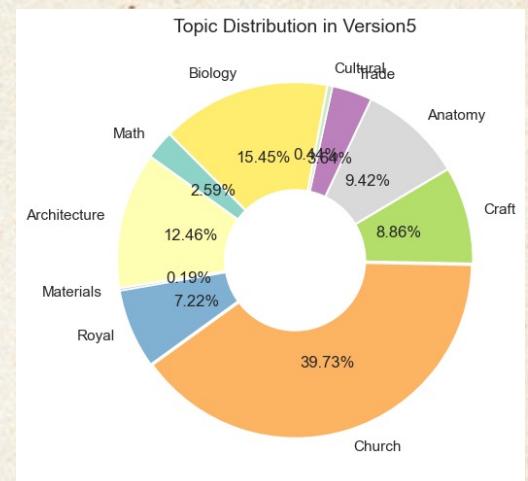
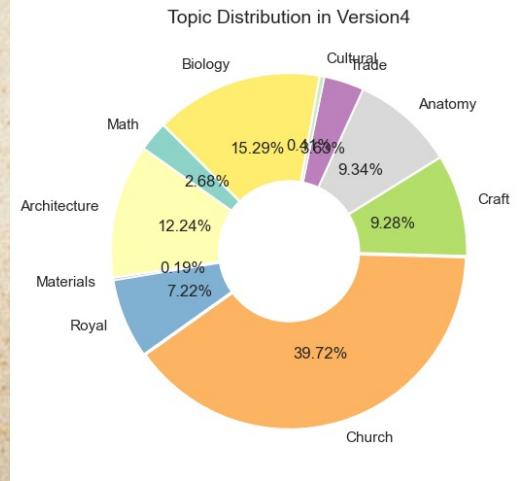
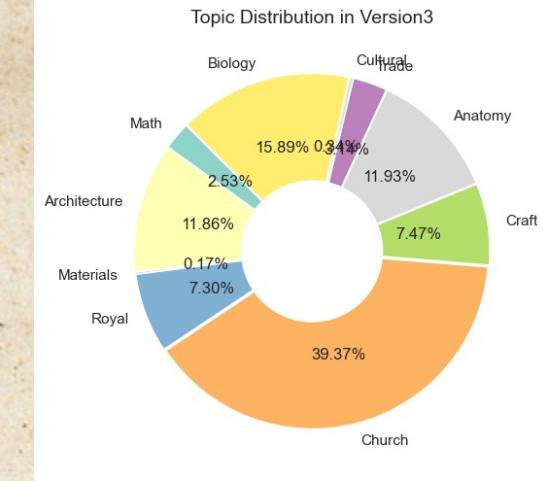
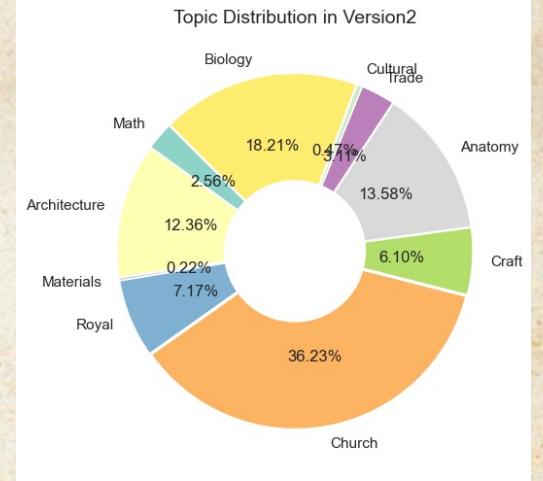
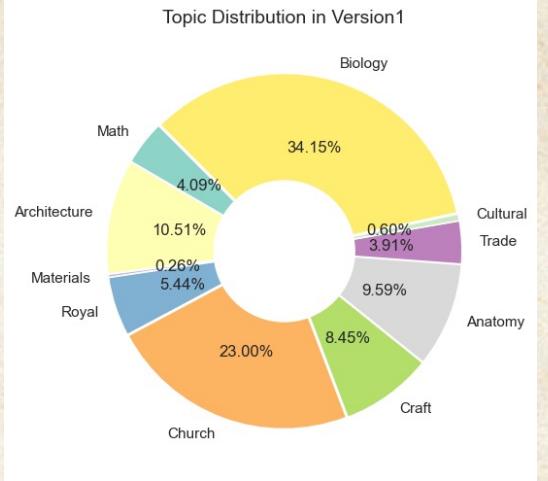
plt.xlabel("Topic Number", fontsize=12)
plt.xticks([x+0.5 for x in range(10)], topic_labels)

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(1824)', 'Year(1830)'],
           rotation=0)

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

## A. Percentage of Visualized Entries





**Percentage of Topics per Edition**

# 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

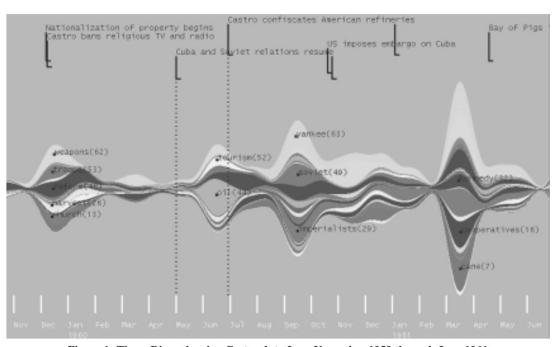
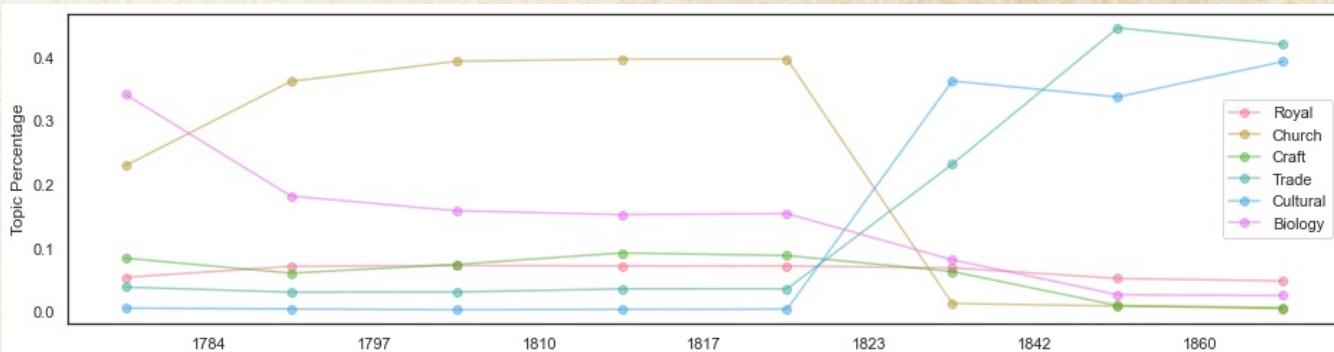


Figure 1: ThemeRiver showing Castro data from November 1959 through June 1961.

### A. Reference

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



### B. "Theme River" in Encyclopedia Britannica

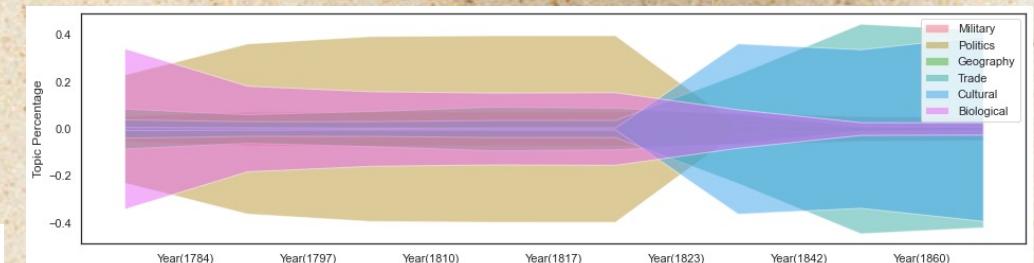
```
sns.set_theme(style="white", palette=sns.color_palette("husl"))

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

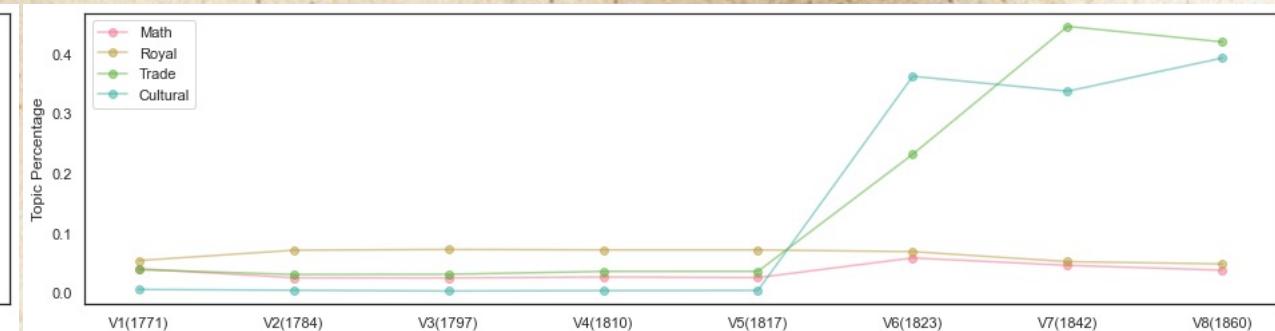
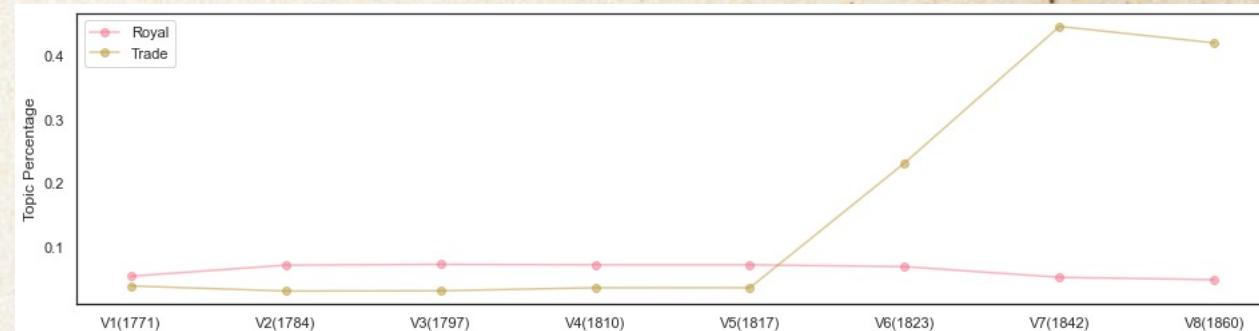
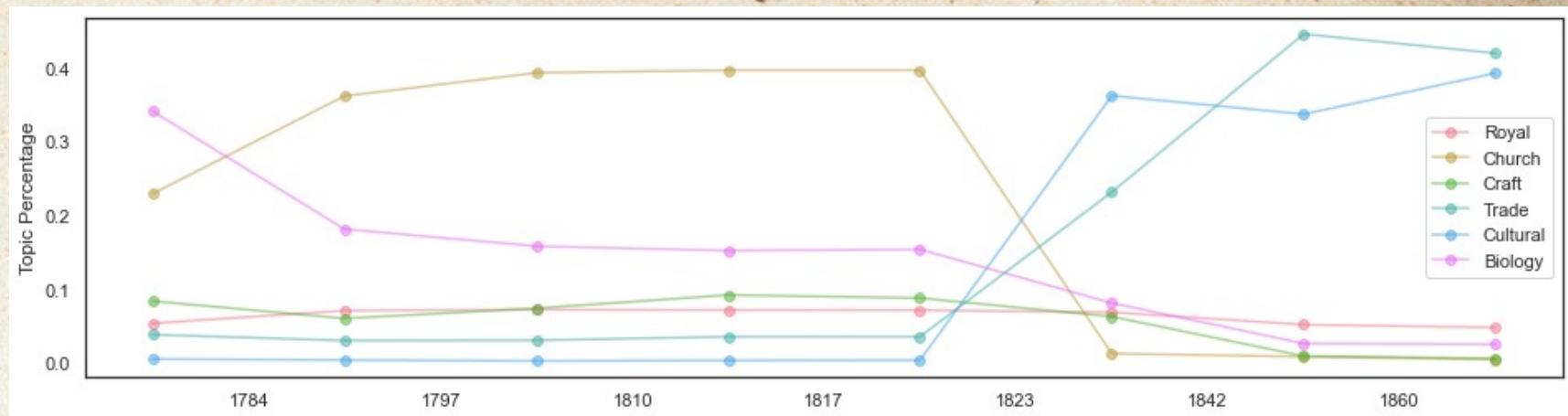
df_topic_pct=df_version_topic.div(df_version_topic.sum(axis=1),axis=0)
plt.figure(figsize=(16,4))
plt.xticks([x+0.5 for x in range(8)],
           [ 'Year(1771)', 'Year(1784)', 'Year(1797)', 'Year(1810)', 'Year(1817)', 'Year(1823)', 'Year(1842)', 'Year(1860)'],
           rotation=0)

for i in [3,4,5,7,8,9]:
    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)
```

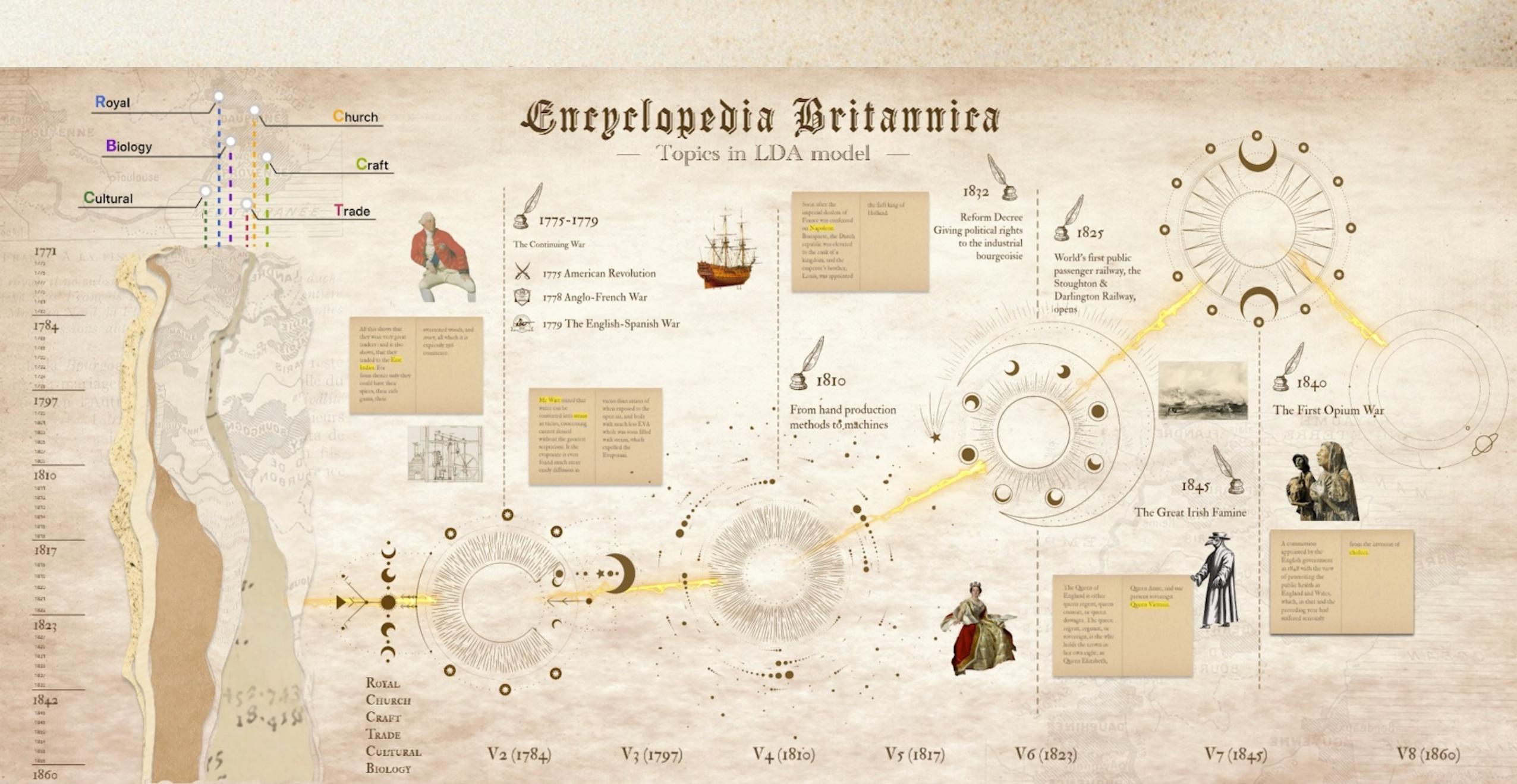


### C. Theme change line chart



## 6. Interactive Board - Arduino and Pressure Sensor

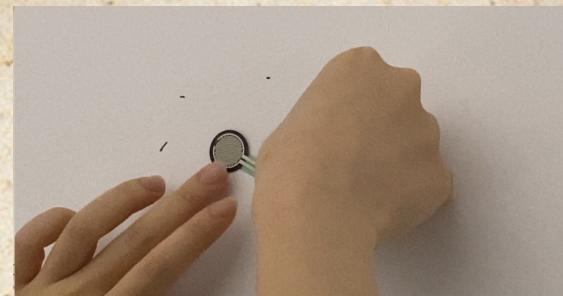




## 7. Modeling Making



01.  
Sticker sensors



02.  
Circuit connections



03.  
Masking sensors

