電気電子情報実験・演習第二: 情報可視化とデータ解析

Final Report

チーム名	Edamame
氏名(学籍番号)	青木悠(03-180397)
	梶浦信勝(03-180409)

構築したシステムの概要，背景を述べてください.

私たちは今回の Capstone Project で、世界で話題のトピックが地図上で一目でわかり、情報の伝播を時間的・空間的に捉え、国ごとの報道の違いを比較できるようなニュース可視化システムを構築しました。現在、地図上の位置情報とニュースが紐づいたアプリケーションが存在せず、世界地図上でニュースを表示するシステムを実装出来たら新たな気付きがあるのではないかと思ったのが実装の動機です。また、その動機から派生した目的を達成する機能として他国のありのままのニュースに気軽に触れやすくするためのキーワード単位のワードクラウドで検索する機能やニュースが地図上で伝播する様子を可視化するための時間軸を調整する機能、ニュースの国による報道の違いを分析するための分析画面などを実装しました。

使用したデータセットの取得先，取得したデータの処理方法を記載してください.

まず「各国で話題になっているキーワード」10 個は、それぞれの国の版の Google News にある、話題のキーワード 10 個を使いました(Google News は約 70 か国で公開されています)。世界中のキーワードすべてを Google Cloud で翻訳し、英語に統一した後、何か国でそのキーワードが話題になっているかをカウントし、それを話題度としました。ここまで行ったのが./storage/csv/以下のファイルで、ワードクラウドのソースとなります。さらに、それぞれの国の 10 個のキーワードの「話題度(の対数)」の総和を、国の重要度としました。このデータを整形したのが./storage/json/以下のファイルで、地図上の国の円のソースとなります。ここまでのすべての処理を、1 時間おきにスケジュール実行し、キーワードに関するデータを蓄積しました。

最後に、取得を行った期間中に 1 回でも話題になったキーワードすべてについて、そのキーワードを話題にした国すべての Google News から RSS の形式で記事一覧(タイトル、ソースのリンク、冒頭部分などが含まれる)を取得しました。タイトルはすべて Google Cloud の翻訳を用い、英語と日本語に訳しました。ここまで行いデータを整形したのが./storage/rss/と./storage/rss_jp 以下のファイルで、記事比較画面のソースになります。

データセットの取得・処理はすべて Python を用いています。

システムで使用している可視化・インタラクションのデザインの根拠(なぜそのようにデザインしたか)を述べてください．また各インタラクションに関して授業中で述べた 7 つの intent のうち，どれに該当するかを記述してください.

あるキーワードとそれを話題にしている国の広がりの関係を見るため、ワードクラウドのキーワードにマウスオーバーするとそれを話題にしている国がハイライトされるようにしました。記事比較のためにさらに国を絞る必要があるので、キーワードをクリックすると、先程ハイライトした国以外は非表示になるようにしました。これは Select/Focus、Filter、Connect に該当します。

また、ある国で話題になっていることが知りたいユーザを想定し、国の円にマウスオーバーするとトピックキーワードのリストが表示されるようにしました。そこでクリックすることで表示が固定され、さらにそこからキーワードを Select できるようにしました。これは Select/Focus、Abstract/Elaborate に該当します。

時間変化に伴うニュースの広がり・話題の変遷を見るため、地図の画面において時刻を変更できるスライドバーを実装しました。これは Explore に該当します。

地図の画面でキーワードをセレクトしている途中で簡単に初期状態に戻せるようにするため、Reset ボタンを作成しました。これは Reconfigure に該当します。

地図の画面・記事比較画面に共通して、国が属する地域によって色を分けています。これは地域ごとの報道の仕方の特性・傾向を見出しやすくするためです。これは Encode に該当します。

記事比較画面において、記事の探索・記事同士の比較を容易にするため、記事のピン留め機能、ソート機能、国のペインの消去機能を実装しました。これは Select/Focus、Filter に該当します。また、キーワードの話題度とその元となった記事の関連を知るため、画面右上に話題度遷移のグラフを表示し、時間範囲を選択するとその範囲にリリースされた記事のみがハイライトされるようにしました。これは Filter、Connect に該当します。

全体を通して、各ボタンやオブジェクトがクリックできることをユーザに伝えるために、マウスオーバーでカーソルの形が変わったり、リンクにアンダーラインが引かれたり、オブジェクトの色が変わったりするようにしました。これは Select/Focus に該当します。

システムを実際に使用してみて得られた興味深い知見を報告してください.

第一に、東南アジアやアフリカで意外にも日本のトピックが話題になっているということが分かりました。一方、日本は他国のニュースをよく報道しているというよりは自国のトピックがよく話題にあがっているという印象でした。

第二に、予想していた通り、対立している国のニュース報道内容の違いが分析できました。例えば、殺害されたサウジアラビアのジャーナリスト Jamal Khashoggi 氏に関しての当事国のサウジアラビアと亡命先のアメリカ、そして第三者のモロッコの記事の違いが挙げられます。亡命先であるアメリカはサウジアラビアの皇太子を批判するニュースを報道するのに対

し、サウジアラビアではアメリカによる批判を批判する内容のニュースが報道されています。一方、モロッコは国連に事件の調査を求めるような客観的に捉えたニュースが報道されています。

もう一つの例として、日本の外国人労働者問題に関して、当事国である日本と多くの労働者の出身国であるベトナム、そして第三者のアメリカやイギリスの記事の違いが挙げられます。日本は外国人受け入れに関する法案が成立した事実を報道しているのに対し、ベトナムではやや日本の外国人受け入れを批判する傾向にあります。一方、アメリカやイギリスはその中間であり、客観的な事実を報道しつつも疑問視しているという記事が報道されています。

第三に、アフリカで技術的なトピックがよく話題になっていることも驚きでした。それに関連して、アフリカのニュースの報道元の新聞社が一部共通していることが分かり、そのためアフリカの国の円が大きくなることが多いということも分かりました。

その他，システムやコード，データセットに関して講義担当者に知らせておくべきこと(例えば，既知のバグやデモを動かす上で注意が必要なことなど)があれば書いてください．システムのインストールに特別な注意や設定が必要な場合にはここにその説明を記載してください.

既知のバグとして、ワードクラウドの描画時間がやや長く、世界地図画面での時間操作の時間がワードクラウドの描画時間より短いとワードクラウドが二重に描画されてしまうというバグが挙げられます。矢印キーによる時間操作機能を撤廃したことで発生しにくくなりましたが根本的な解決には至っていません。

Google Translation のバグで、Cookie が有効な状態では表示されたリンクの先のページ全体の翻訳がうまくいかないことがあります。プライベートブラウズを用いることでこの問題を回避できます。

システムの主な画面は index.html ファイルを開くことによって実行できます。index.html 内で compare ボタンを押すことにより、記事比較画面である comp.html が実行されます。

データ収集・データセット作成のために用いたプログラム(translate_keyword.py、translate_article_rss.py)は Google Cloud の API を使用します。これらを動かすにはライブラリのインストールと有効なアカウント情報を記した json ファイルが必要です。

また、Slack でも予め相談しましたがワードクラウドの描画に関して d3.layout.cloud.js というライブラリを使用しています。