# ゼミナールの特徴をタグクラウドで視覚化するWebサイトの作成

斎藤ーゼミナール
1122065　佐藤　佑起

# 背景

- 情報メディア学科のゼミナール（以下ゼミ）の数が多い
- シラバスで提示されているイメージと実際のゼミの活動があっていない場合がある。
- 情報大学内のWEBサイトでは継続的にゼミを紹介しているサイトがない。

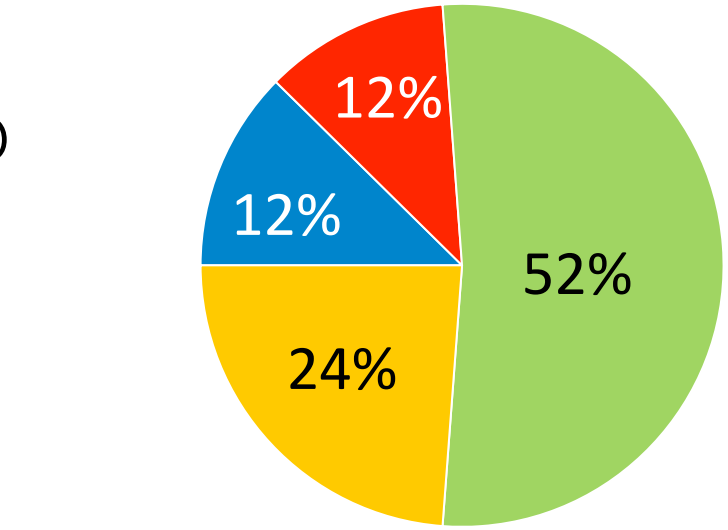


情報メディア学科に所属する3年生以上（112人）に
アンケート調査を行う

# 結果

ゼミを調査する期間は2年生の
後半である割合が多い

どのゼミに入ろうか情報が
少なく迷う




ゼミについて2年時のいつから調査を始めたか

卒業論文報告書をタグクラウド化し、

各ゼミの特色が一目でわかるWeb

# タグクラウドについて

## タグクラウドとは・・・

- コンテンツやURLなどを**タグ**と呼ばれる、**単語**や**短いフレーズ**で分類しているサービスやサイトにおいて、利用されている**タグを一覧表示**したもの。また、そのような表示方法や表示するためのツールなどのこと。

（※　IT用語辞典e-Wordsより引用）


※collosopns.dotimpac.toより引用


※goowikipedia記事検索より引用

# 報告書からキーワード（タグ）を抽出

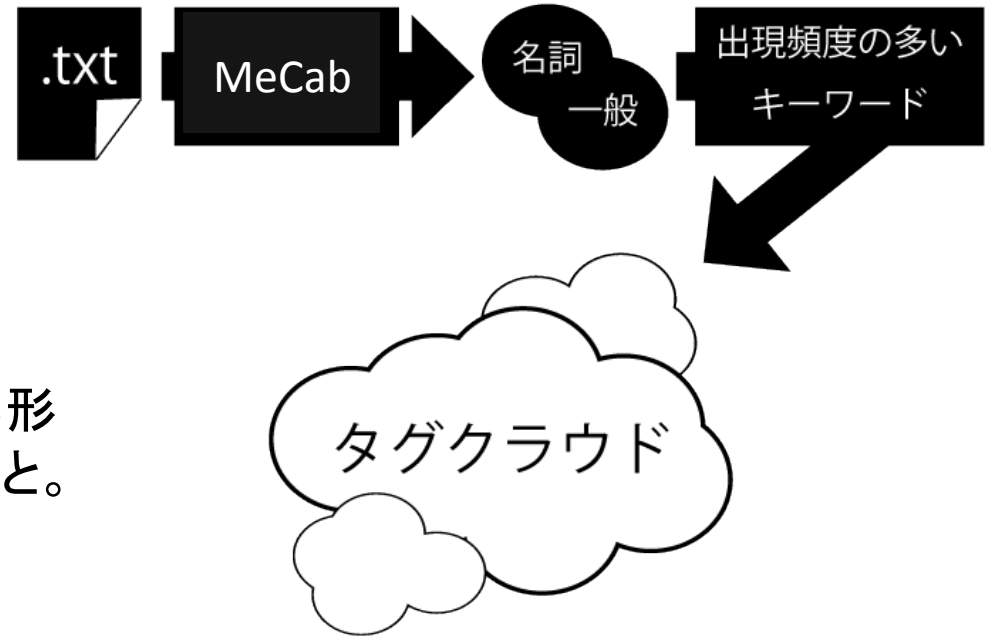


1．報告書を形態素解析する

2．名詞と一般が含まれるキーワードを抽出

3．出現頻度の高いキーワードを抽出

・形態素解析ソフト**MeCab**を使用




**※形態素解析とは？**
文章の意味を担う最小の単位である形態素ごとにコンピュータで分割すること。
（「自然言語処理-基礎と応用-」より引用）

# キーワード選別方法

- TF-IDF法を使用

- TF（TermFrequency）
　それぞれの単語の文章内での出現頻度を表し、出現するほど単語としての重みをつける。

＜計算式＞

$$\text{単語[X]のTF値} = \frac{\text{単語[X]の文書内での出現回数}}{\text{文書内の全単語の出現回数の和}}$$

TF値が0.006より上の単語を抽出する。

# キーワード選別方法

- IDF（InverseDocumentFrequency）

　単語がどれだけの**文書集内で共通して使われている**かを表す指標。その中で、出現頻度が高くなれば重要度が低くなる。

※本研究では擬似的なIDF法を使用する。

＜方法＞

　2014年卒業研究報告集からすべての報告書で出現している単語をストップワードとして登録し抽出しないようにする。

＜ストップワード例＞

- 図
- 例
- ゼミ
- 人
- 情報
- 学科

# キーワード選別方法



- 出現頻度は高いが特徴としては不適切等の不要単語をストップワードとして登録する（全７７ワード）
  - 開発環境
  - 制作環境
  - http
  - www
  - 目的
  - 筆者
  - 理由

等

サイトからテキストファイルをアップロード

↓

キーワード抽出＆選別

↓

データベース（MySQL）に登録

- タグクラウド表示方法

フォントサイズはhtmlの見出しを使用

見出しサイズと出現頻度順の関係性

- <h1> ＝ 1番（配置位置：中央）
- <h2> ＝ 2番～3番
- <h3> ＝ 4番～7番
- <h4> ＝ 8番～12番
- <h5> ＝ 13番～17番
- <h6> ＝ 18番～20番

| keyword | tf | count |
|---|---|---|
| 写真 | 55 | 6 |
| 効果 | 13 | 5 |
| サイト | 25 | 5 |
| アンケート | 27 | 5 |
| 斎藤 | 8 | 5 |
| 被験者 | 15 | 4 |
| Webサイト | 18 | 4 |
| ゲーム | 23 | 4 |
| コンテンツ | 11 | 4 |
| メディア | 6 | 4 |
| 方法 | 10 | 4 |
| 種類 | 8 | 4 |
| キャラクター | 11 | 3 |

※CSSでボックスのサイズと座標を指定し、各ボックスにトップワード以外をランダムで配置する

## 使用技術

| | | |
|---|---|---|
| **開発環境** | Ubuntu14.04LTS<br>HTML5<br>CSS3<br>Jquery | Apache　2.4.7<br>PHP　5.5.9<br>Pear |
| **使用技術** | MeCab0.98<br>MeCab ipadic(MeCab用辞書) | MySQL |
| **使用ソフト** | VMware player<br>Ubuntuテキストエディタ | Adobe　Illustrator CS6 |

- 2年生7人に利用した感想を調査（1月23日）
  - 7、8割はイメージ通りの結果であった
  - イメージとは違う単語が出てきた
    例）斎藤一ゼミでは「観光」　等
  - ゼミを選ぶ参考になった
  - 同じような特徴をもつゼミでもほかの特徴によりゼミ決定の際に参考になった

# まとめ

- 形態素解析を使用し、報告書から特徴となるキーワードの抽出に成功
  ※計18ゼミを解析終了（ストップワード調整終了は2ゼミ）
- タグクラウドの表示
- キーワード選別方法の実装


大島慶太朗ゼミを解析した結果

## 今後の課題

- 更に高精度なキーワード抽出＆選別方法の模索
- ストップワードに登録する単語の考察
- シラバスに表記されているゼミの特徴の中にはサ変接続のキーワードもあるためサ変接続を含んでのさらなる調整の必要性


サ変接続をすべて含んだ際の抽出結果