各種データ分析ツールの紹介

坂田・森研究室
特任研究員　榊　剛史

本講義の目標

プログラムを書かずにデータ分析を行う
※主に学術俯瞰システムの分析出力

- ネットワーク分析ツール
  - GePhi
  - Cytoscape
  - NodeXL
- テキストマイニングツール
  - TTM
  - ExcelTTM
  - **Googleスプレッドシート上で形態素解析**

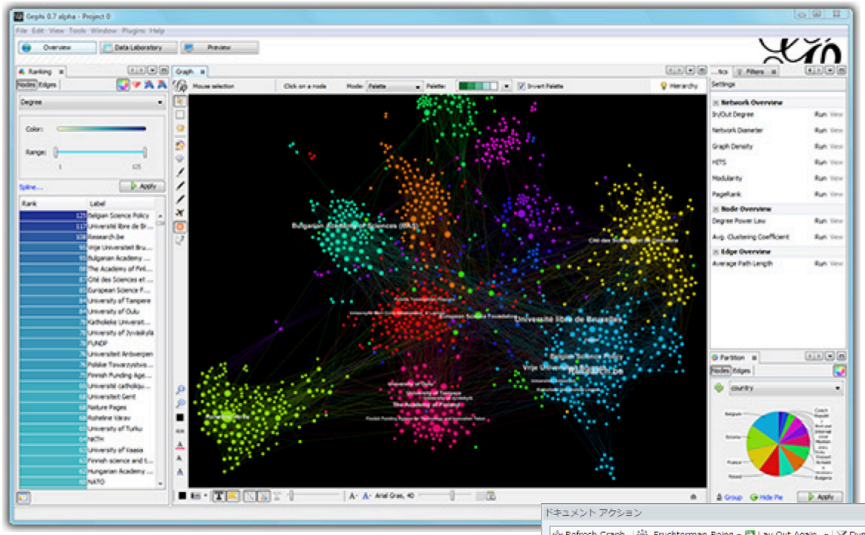
ツールにデータをインポートする方法をメインに説明

ネットワーク分析の基礎

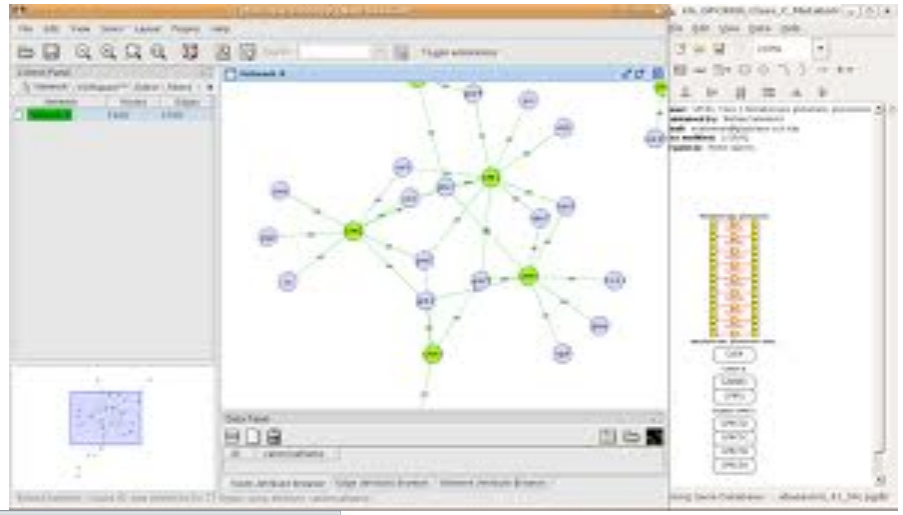
- 社会ネットワーク分析とは
  - 現実世界に存在する巨大で複雑なネットワークの性質について分析すること
  - 事例:「蛍の木の発光同期現象」
    - https://www.youtube.com/watch?v=Ls6jJnJ2CuQ
  - 分析事例
    - 論文ネットワーク内で重要度の高い論文を発見する
      - 各論文の中心性を算出
    - 論文ネットワークをグルーピングする
      - クラスタリング手法の適用

ネットワーク分析ツールの紹介

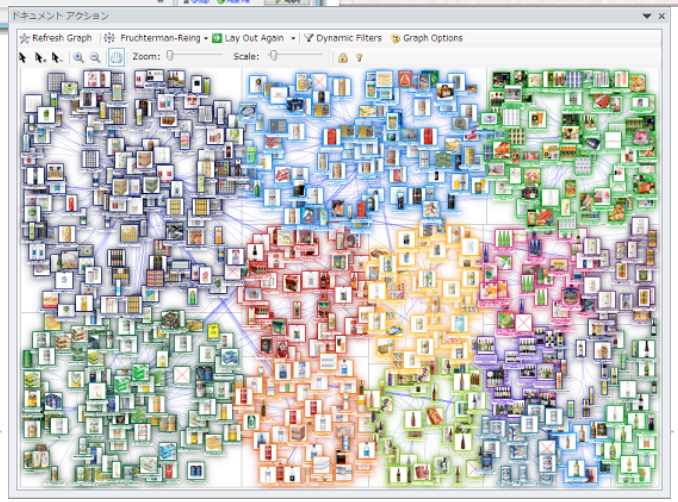
GePhi



Cytoscape



NodeXL



ネットワーク分析ツール比較

	GePhi	Cytoscape	NodeXL
歴史	浅い	古い	中程度
使い勝手	普通	普通	容易 (Excelベース)
機能性	多機能	多機能	限定的
拡張性	高い	高い	低い
大規模データ	数千程度	数万程度	千程度
OS	Win/Mac/ Linux	Win/Mac/ Linux	Win

GePhi

- 特徴:
  - 多様なネットワーク形式に対応
  - クラスタリングや中心性計算など様々な分析が可能
  - 他のユーザが開発した追加プラグインを利用することも可能

- サポートしているデータ形式
  - CSVファイル
  - GEXF
  - Pajek NET
  - GraphViz DOT
  - UCINET DL

GePhiの使用方法

- CSV形式
  - nodeファイルとedgeファイルの2つのファイルにより構成される
  - nodeファイルには各ノードのIDやラベルを記述する
  - edgeファイルには各エッジのノードペア情報やラベル、有向・無向などの情報を記述する

GePhiの使用方法

- エッジファイル

ノード1	ノード2	有向性	エッジID	重さ

	A	B	C	D	E	F
1	Source	Target	Type	Id	Weight	Average Degree
2	1	3	Undirected	1	1	1
3	2	10	Undirected	8	1	1
4	3	4	Undirected	2	1	1
5	4	5	Undirected	3	1	1
6	5	10	Undirected	4	1	1

  - SourceとTarget列が最低限あればOK
  - Excelで編集→別名で保存「CSV形式」で作成・編集可能

GePhiの使用方法

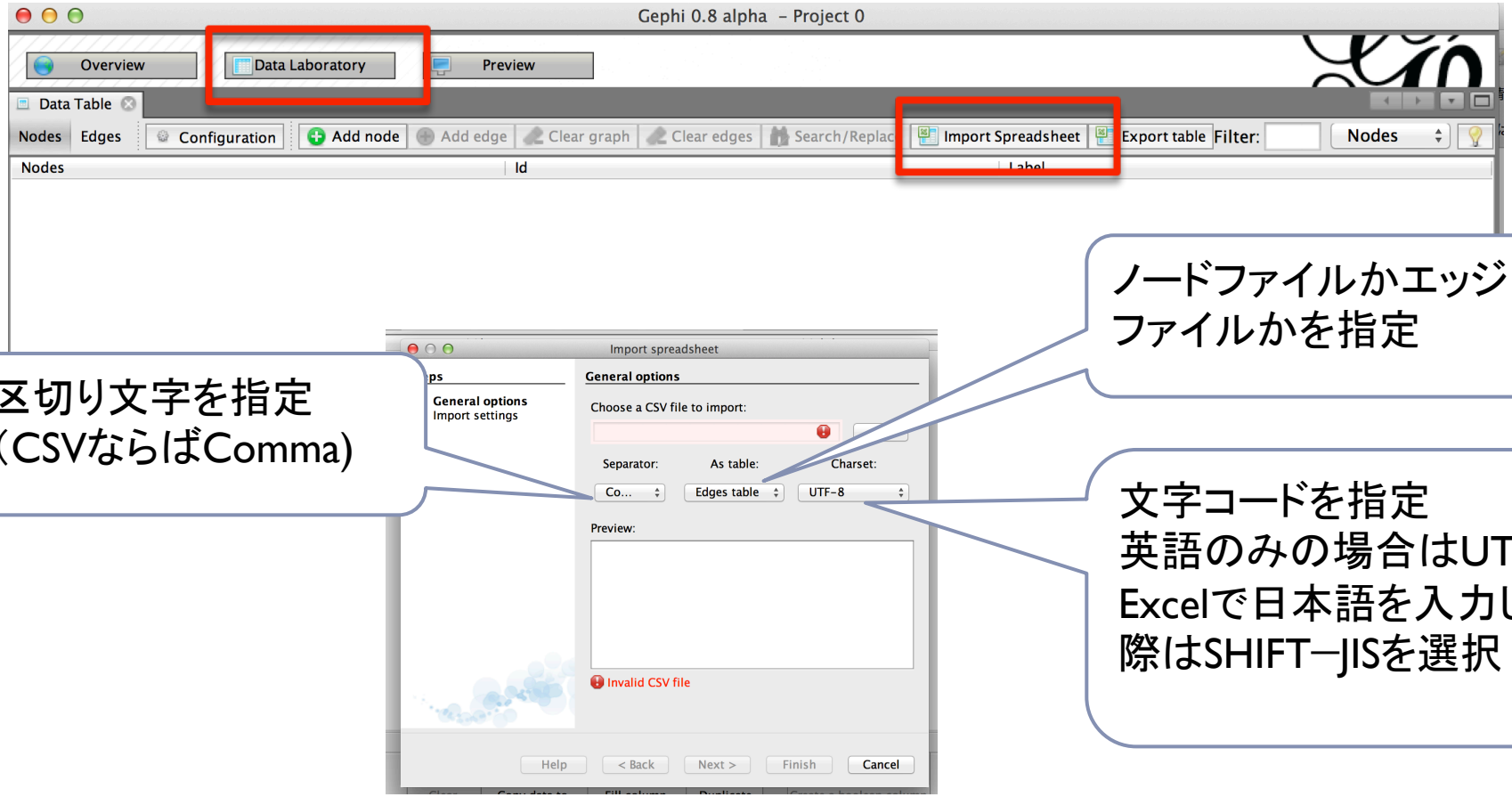
- ノードファイル

ノードID	ラベル名

	A	B	C
1	Id	Label	SquareFootage
2	1	Office	224
3	2	Kitchen	230
4	3	Clothes Closet	45
5	4	Storage Closet	56

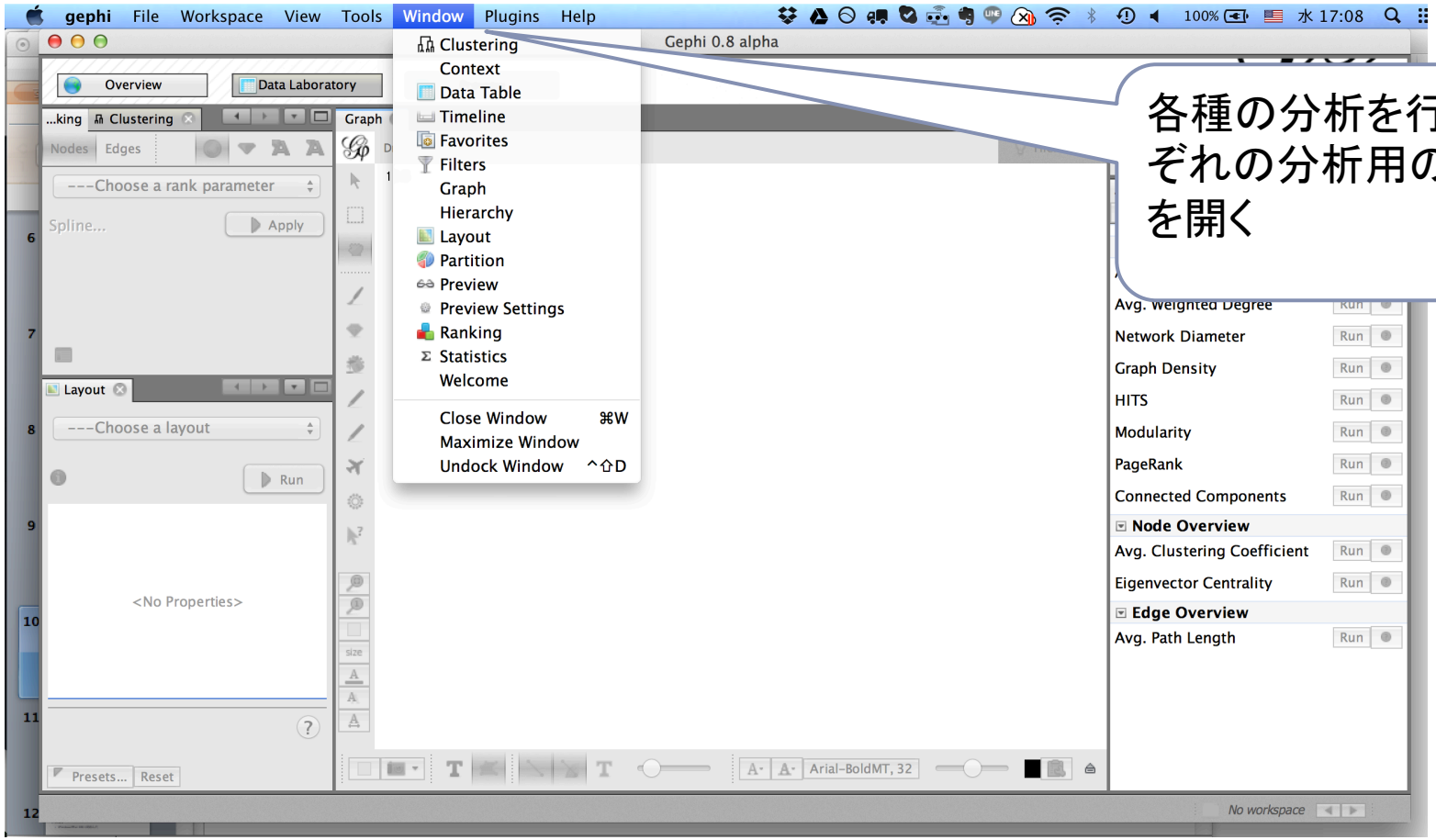
  - idとLabel列が最低限あればOK
  - Excelで編集→別名で保存「CSV形式」で作成・編集可能

GePhiの使用方法



エッジファイルとノードファイルをそれぞれ、計2回インポートを行う

GePhiの使用方法



https://gephi.github.io/users/quick-start/

テキストマイニングの基礎

- テキストマイニングとは
  - 通常の文章からなるデータを単語や文節で区切り、それらの出現の頻度や共出現の相関、出現傾向、時系列などを解析することで有用な情報を取り出す、テキストデータの分析方法

- 分析事例
  - ある論文クラスタの特徴を表す語を抽出する
  - 内容的に類似した論文や論文クラスタを発見する
  - 類似した文クラスタと特許クラスタを紐付ける

テキストマイニングの基礎

- テキストマイニングの問題点
  - テキストはそのままだと数理的に分析することが困難

  →数理的なモデルで表現してやればよい

- ベクトル空間モデル
  - 文書情報の数理モデルのひとつ で、文書をベクトル空間内の「点」と表現　　$\mathbf{d} = (t_1, t_2, \cdots, t_n)$
  - 通常、文書は大量に存在するので行列表現となる
    - 単語ー文書行列

単語

$$\mathbf{d} = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ t_{m1} & t_{m2} & \cdots & t_{mn} \end{bmatrix}$$ 文書

テキストマイニングの基礎

- テキストマイニングツール
  - 文書を単語に分割し、文書ー単語ベクトルが作れれば良い
- Tiny Text Miner　http://mtmr.jp/ttm/
  - Windows/Mac OSに対応したTTMツール
  - 機能は少ないが使いやすい
- ExcelTTM　http://mtmr.jp/excelttm/
  - Excel上で実行可能なTTM
  - Windows版/Mac OSに対応しているが、Mac OSではExcel2010が必要
- Google Spread Sheet + Yahoo API
  - http://shirayuca.github.io/blog/2014/08/04/yahoo.html
  - Google Spreadsheet上で単語集計を行うソースコード
  - ツールとして整備されているわけでは無いので注意

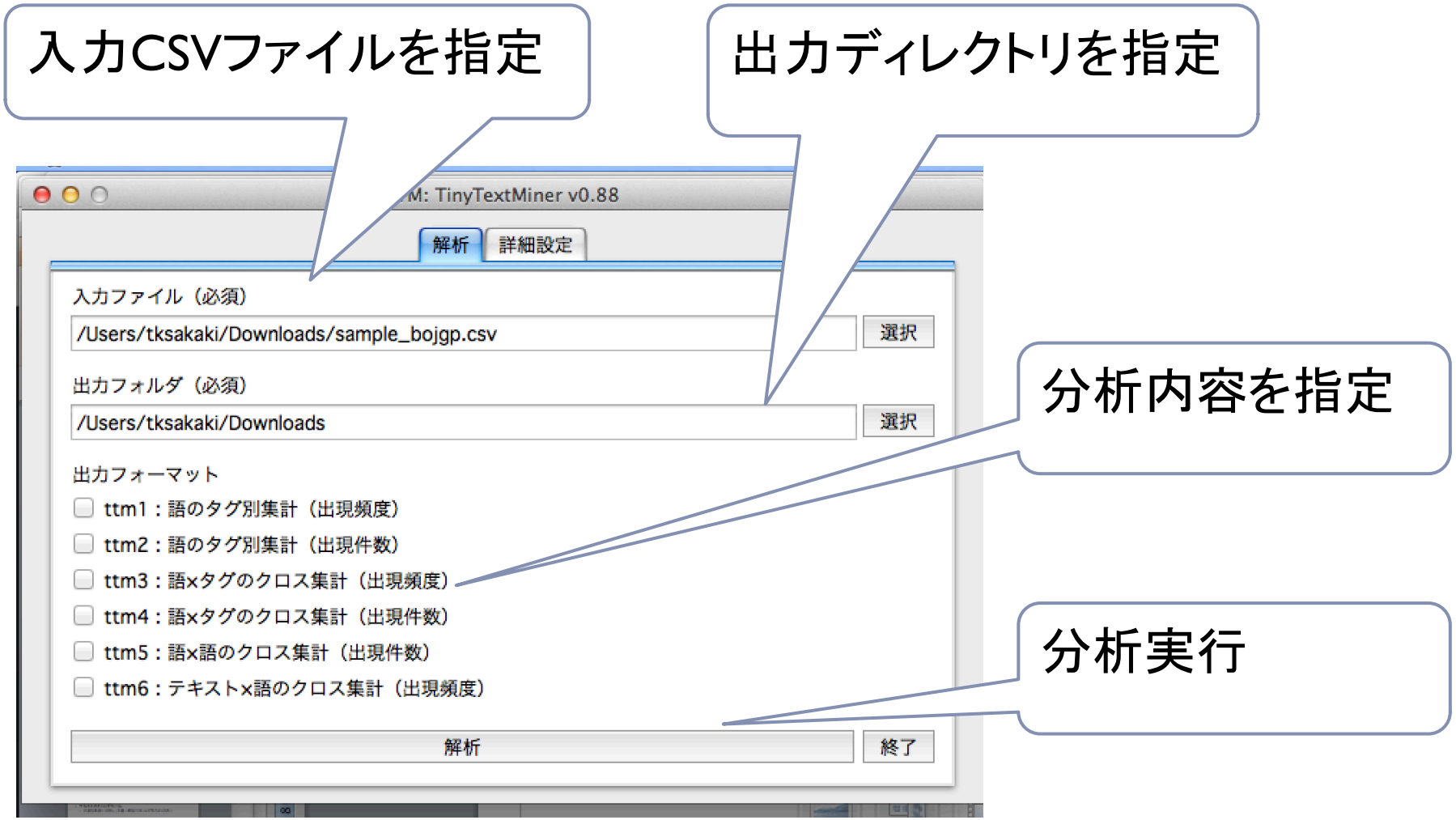
Tiny Text Minerの使い方

- 機能
  - タグが付与されたテキストに対して、単語の集計、ラベルの集計などが可能

- 入力形式：CSV形式
  - タグ、文書で構成されるN行×2列のCSVファイル

タグ	文書
200612	わが国の景気は、緩やかに拡大している。
200612	公共投資は減少傾向にあるが、輸出は増加を続けている。また、企業収益が高水準を続け
200612	先行きについても、景気は緩やかな拡大を続けるとみられる。
200612	すなわち、輸出は、海外経済の拡大を背景に、増加を続けていくとみられる。また、国内民
200612	物価の現状をみると、国内企業物価は、原油価格の反落が影響し、足もとでは3か月前比
200612	物価の先行きについて、国内企業物価は、原油価格反落の影響がなお残ることから、目先
200612	金融面をみると、企業金融を巡る環境は、緩和的な状態にある。CP・社債の発行環境は良
200703	わが国の景気は、緩やかに拡大している。
200703	公共投資は、足もと幾分増加しているが、基調としては減少している。一方、輸出は増加を
200703	先行きについても、景気は緩やかな拡大を続けるとみられる。

Tiny Text Minerの使い方



Tiny Text Minerの使い方

各タグごとに単語の出現頻度を集計

タグ＝文書と考えれば、容易に単語-文書ベクトルを構築することができる

シート

	A	B	C	D	E
1	タグ	語	品詞	出現頻度	
2	200612	続け	動詞	7	
3	200612	増加	動詞	7	
4	200612	み	動詞	5	
5	200612	水準	名詞	5	
6	200612	緩やか	形容詞	5	
7	200612	増加|し	動詞	5	
8	200612	前年|比	名詞	3	
9	200612	み|る	動詞	3	
10	200612	中	名詞	3	
11	200612	あ|る	動詞	3	
12	200612	な|っ	動詞	3	
13	200612	前月	名詞	3	
14	200612	推移|し	動詞	3	
15	200612	背景	名詞	2	
16	200612	拡大	動詞	2	
17	200612	公共|投資	名詞	2	
18	200612	増加|基調	名詞	2	
19	200612	き	動詞	2	
20	200612	物価	名詞	2	

Google Spread Sheet + Yahoo API

- サイトの手順に従ってデモを行います
  - http://shirayuca.github.io/blog/2014/08/04/yahoo.html

- 手順
  - Yahooデベロッパーズサイトに登録し、AppIDを取得
  - Google Spreadsheetを作成
  - Spreadsheet上にスクリプトを追加
  - テスト実行

まとめ

データ分析を行うためのツールを紹介

基本アプローチ

・ExcelでCSVファイルを作成・変数する
・各ツールにCSVを読み込ませる
（Excelのマクロを用いてExcel上で分析する

ネットワーク分析ツールについては、ウェブ上に知見が多数存在しているので、自学自習すること