メタデータ自動抽出の技術検証

➢ コミュニティチャンネル、多チャンネルのコンテンツ流通やEPG用のメタデータ作成に伴う、労力削減を目的とした番組メタデータ自動抽出の実現性を調査

番組データ → 抽出方法 → メタデータ項目 → 評価

- アニメ・特撮 **105件**
- スポーツ **104件**
- 海外ドラマ **100件**
- 邦画 **107件**

→ ルールベース

→ 形態素解析

- 「監督等」の後を抽出 → **スタッフ-氏名** → 正解率 **約84%**
- 「xx分」を抽出 → **尺** → 正解率 **約81%**
- 「アカデミー賞等」を抽出 → **賞名** → 正解率 **82%**
- 「/アメリカ等」を抽出 → **制作国** → 正解率 **約97%**
- TF-IDFスコア算出 → **検索キーワード** → **28,428個** 単語抽出
- Word2Vec合致度算出 → **ジャンル** → ・スコアトップ **約13%** ・スコアトップ5内 **約61%** ・スコアトップ10内 **約83%**

TF-IDFとは

・TF(Term Frequency):
単語ごとの文書内での出現割合

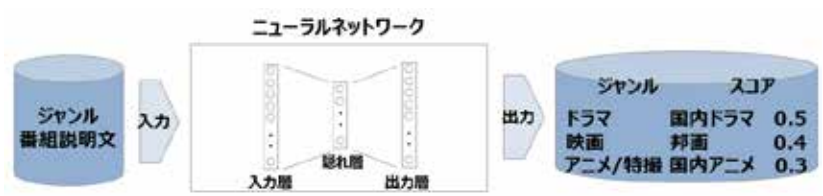
・IDF(Inverse Document Frequency):
ある単語が出現する文書の全文書での出現割合

■番組の「検索キーワード」のスコアリングは、番組説明文より単語を抽出し、単語ごとTF-IDFによりスコア算出

Word2Vecとは

●今回の入出力データ

ジャンルと番組説明文を解析し、意味ベクトル(word2vec)の類似度の高いジャンルを抽出



JLabs
Japan Cable Laboratories