DojinHakusho User Manual

DojinHakushoをダウンロード頂きありがとうございます.

本ツールは, DLsiteで活躍するクリエイターの情報を収集・解析するフリーソフトです. 特に同人音声で活動するサークル様・声優様の作品を解析対象とした利用を想定しています.

本ファイルにはツールのインストール方法と基本的な使用方法が記載されます. 内容を熟読されたうえでご利用いただきますようお願い申し上げます.

> 注意事項:
>
> ツールも解説も真面目ですが, 取り扱う分野の性質上一部アダルトな内容が含まれます. 18歳未満の方は閲覧をお控えください.

利用規約

本ツールはDLsiteを利用するクリエイター様に対する支援を目的に開発されたツールです. 従って, 本ツールによって得られるデータを, その目的から大きく逸脱して使用することはお控えください. (例えば, データを用いた対立煽りや中傷行為などがこれに該当します.)

また, 本ツールはDLsite様にアクセスしHTMLソースを取得することで動作するクローラーです. 従って, 本ツールの利用は常識の範囲内で行い, サーバ管理者から攻撃と見做されるような行為はお控えください. (つまり, robots.txtを守ってください.)

上記規約に従う限り, 本ツールは完全にフリーです. リンク, 商用利用, 転載, 改変, 二次配布などは自由に行っていただいて構いませんし, 連絡の必要もありません. なお, 製作者はその結果生じたいかなる事態にも責任を負いかねます.

インストール

環境構築

python3および必要なライブラリをインストールしてください.

必要な外部ライブラリは, クロールのためのrequests, 基本的な統計調査, 作図, 科学計算のための networkx, matplotlib, pandas, numpy, scipy, japanize-matplotlib, 自然言語処理のためのMeCab, CaboChaです. インストールに別途pipやら何やらが必要になります. 導入方法に関しては, 私の解説よりも遥かに分かりやすいものが出回っているので適当に検索してください.

インストール

DojinHakushoの頒布場所へアクセスし, ツールを任意のディレクトリにダウンロードしてください. そうしたらzipが出てきますので, 任意のディレクトリに解答してください.

使用方法

まず『main.py』を適当なエディタで開き，この行に調べたいサークル名と声優名を入力してください．名前は必ずDLsiteのクリエイタータグ名と一致させる必要があります．検索対象となるデータは，変数で設定されたサークル様作品と声優様出演作品の積集合です．どちらも不問の場合は空欄で構いませんが，両方を空欄にすると実行時間が長くなるのでオススメしません．

なお，今回はサークル名を空欄，声優名を砂糖しお様に設定したと仮定して解説を行います．

```
# ここに調べたいサークルの名前を入力する．
# 名前はDLsiteのサークル名と一致させる必要がある
# サークル不問の場合は空欄にすること
circle_name = ""

# ここに調べたい声優の名前を入力する．
# 名前はDLsiteのクリエイタータグと一致させる必要がある．
# 声優不問の場合は空欄にすること
creator_name = "\"砂糖しお\""
```

データを取得するにはコマンド上でインストール先のディレクトリに移動し，以下のコマンドを入力します．

```
> cd DojinHakusho
> python3 main.py
```

上記の命令でツールを実行すると，このような実行結果が得られるかもしれません．なお，本ツールは検索範囲を**「販売中の男性向けオーディオファイル」**に絞っています．

```
> 検索条件：
>  サークル名　　："" 
>  声優名　　："砂糖しお"
>  販売状況　　："販売中"
>  対象性別　　："男性向け"
>  ファイル形式　："オーディオファイル"

> 探索終了までお待ちください…

> 探索終了！
> 作品数　　：　127 [個]
> 実行時間　：　23.5 [sec]
```

実行が終了したため，得られた生データ（.csvファイル）を確認してみましょう．ツールが問題なく終了した場合，生データは/data/raw_data.csvに生成されるはずです．

（ファイルが文字化けする際は，一度テキストエディタで開き，文字コードをANSIに変更してください．）

データが得られたら，お好きなスクリプトを実行してグラフを描画したり統計量を得たりしてみましょう．今回は全活動期間の要約統計量を得るために，以下のコマンドを用いて『getDescribe.py』を実行します．

```
> cd script
> python3 getDescribe.py
```

すると，以下のような結果を得ることができるかもしれません．基本的なデータの読み方は次項で解説します．

```
             price        sales
count   127.000000   127.000000
mean   1063.622047   986.755906
std     474.051550  1564.399041
min     110.000000    22.000000
25%     880.000000   229.000000
50%    1100.000000   460.000000
75%    1100.000000   921.000000
max    3520.000000  8972.000000
```

以上がDojinHakushoの基本的な使い方の流れです. 同様に, 適当にいくつかスクリプトを走らせて遊んでみてもよいでしょう.

各スクリプトの解説

本項では, /scriptフォルダに存在するスクリプトを走らせて, 得られる統計量やグラフに対する簡単な解説を行います.

なお, 本項で使用するデータはいずれも2020年10月22日に取得された砂糖しお様のデータである点に留意してください.

> 製作者の一言:
>
> この部分には製作者のゆるい感想や雑記が記述されます. 定量的な評価のみを求める場合は読み飛ばしていただいて構いません.

allinone.py

下に示すスクリプト全て(ただし[試作]と付いているものは除く)を一気に回します. 面倒なときはこれを使ってください.

getDescribe.py

とりあえず得たデータの性質を手っ取り早く得たいので, このスクリプトを用いて要約統計量を入手します.

```
             price        sales
count   127.000000   127.000000
mean   1063.622047   986.755906
std     474.051550  1564.399041
min     110.000000    22.000000
25%     880.000000   229.000000
50%    1100.000000   460.000000
75%    1100.000000   921.000000
max    3520.000000  8972.000000
```

このログは, /log/getDescribe.txtに生成されます.

まずは, 上記要約統計量についての解説を行います.

countはデータの個数, meanは平均値, stdは標準偏差です. min,maxはそれぞれ最小値と最大値を示します. 25%, 50%, 75%はそれぞれ第一, 第二, 第三四分位数を指します. 今回の解説では, 第二四分位数, つまり中央値をベンチマークとして用います. これは, 同人音声界隈は一部の大手と言われるサークルが大量の売上本数を出す傾向にあり, 平均値が実情と剥離すると判断したためです.

データから, 出演作の50%が880円から1100円の範囲でリリースされていること, 出演作の50%が460本以上の売上本数を出していることを読み取ることができます.

仮に中央値である1100円でリリースし460本を売り上げた場合, 総売上は約50万円, そこからDLsiteの手数料（多分4割ぐらい？）を引くと大体30万円程度, そこから声優様・イラストレータ様の報酬やら制作費やら税金やらを引いた額がサークル様の取り分となります.

もちろんこの試算はセールやクーポンを考慮していないので, 実際の値とは剥離があります.

> 製作者の一言:
>
> 勘違いを防ぐために言っておきますが, これらの統計量は『高い=凄い』『低い=駄目』という意味の指標ではありません. 例えば, 新規サークルだろうが大量に出まくる様な方は, 必然的に中央値や平均値が下がる傾向にあります. 更に, 当然のことながら演技力や喉の強さ等はこれら統計量から得ることができません.

getDescribe_by_years.py

先程のデータは活動期間全てを集計したものなので, 年毎のデータは分かりませんでした. このスクリプトは年毎の要約統計量を導出するために使用されます. (今回は2019年出演作のデータのみを抽出し, 以下に示します.)

```
2019年:
              price        sales
count    49.000000    49.000000
mean   1032.653061  1132.408163
std     462.793940  1769.313555
min     440.000000    57.000000
25%     770.000000   275.000000
50%    1100.000000   460.000000
75%    1100.000000   977.000000
max    3300.000000  7540.000000
```

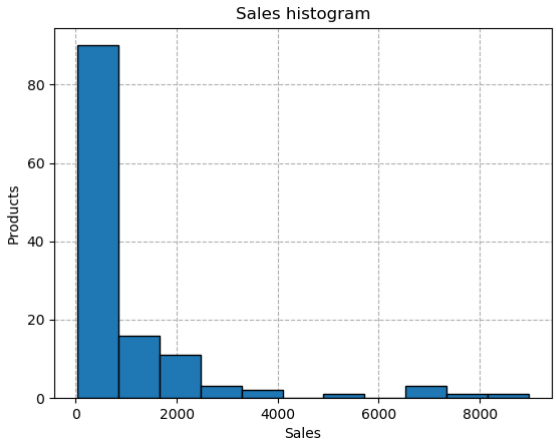
このログは, /log/getDescribe_by_year.txtに生成されます.

出演数は49本. 単純計算で7日に1本のペースで新作が出ていると言えます. これは2019年における出演数ランキング第10位に相当するらしいです. 中央値は値段, 売上ともに上と変わりません.

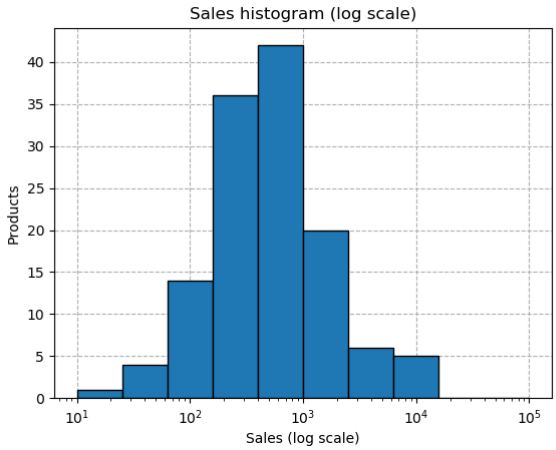
> 製作者の一言:
>
> フリーで活動されている声優様って大体は個人事業主ですよね. つまり, これだけ収録してる傍らで, スケジュール管理から事務処理から対人折衝から技術的な事から税金周りまで全部ご自分で行ってるんですよね. 個人的な感想を言うと凄いとしか言いようがないです.

getHistogram.py

このスクリプトは, 全活動期間を通した売上のヒストグラムを入手します. binの個数はデータ総数の平方根で与えられます. 画像は/fig/Histogram.pngに生成されます.



要約統計量からある程度予想はできていましたが, 作品の大半は売上数<=10^3です. 見やすくするために, 同じデータを対数スケールで表してみます. 対数スケールのヒストグラムは, スクリプト実行時に同時に/fig/Histogram_log.pngに生成されます.



まぁそこそこ綺麗になったのではなかろうか.

先行研究からも売上がlogスケールに従うってのは予想通りです. 取り立てて解説するほどのグラフでもないです.

getPriceSalesPlot.py

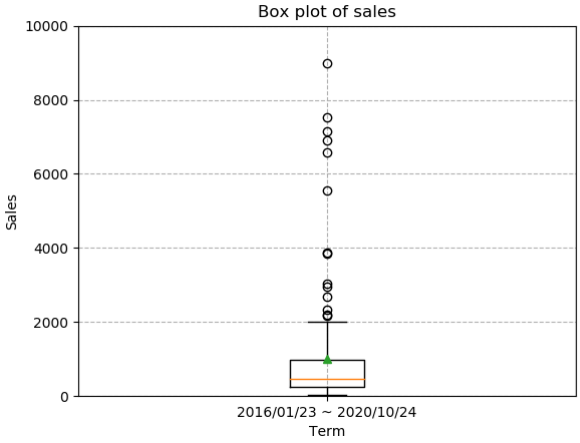
このスクリプトは売上と価格の関係をプロットします. 画像は/fig/getPriceSalesPlot.pngに生成されます.

回帰分析は一応行いましたが, あまり面白いものが得られなかったので示しません. 基本的には1000円前後の価格設定が多いですが, 2000円近い価格設定でもそれなりの売上を出している作品はあるようです. 極端に価格の低い作品は売上を伸ばす傾向にあります. これは他声優様の出演作でも同様の傾向が見られました.

DLsite様が頻繁にセールを行っている他, 最近はサークル様が新作発売日に割引やクーポン発行を行うことも多いので, 正確に実態を表したものではないことに留意ください.

getBoxPlot.py

このスクリプトは全活動期間の売上げから箱ひげ図を生成します.



matplotlibはデフォルトで箱の1.5倍以上の距離にある値を外れ値として扱います. この解説では便宜上, 売上上位25%の作品を『相対的人気作』と定義し, 更に上にある外れ値の作品を『相対的ヒット作』と定義します. この箱ひげ図から言えば, 売上本数第三四分位数以上を相対的人気作, 売上2000本以上を相対的ヒット作と定義できます. それぞれ上位3作を簡単に紹介します.

- 相対的ヒット作（2000本～）

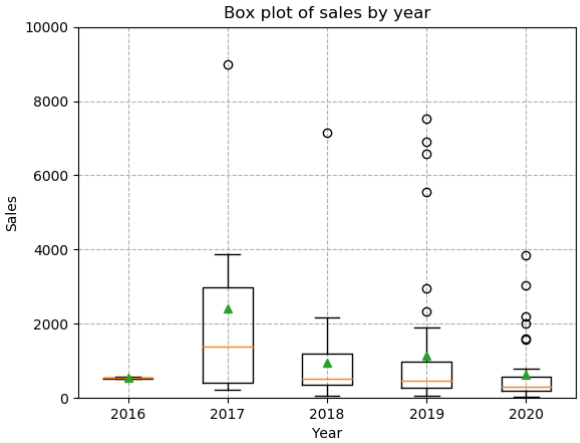
タイトル	タグ構成	売上本数
音声で手コキ 射精はゲームオーバー★快楽我慢ゲーム「早漏遺伝子滅亡計画1」～優しいお姉さんからの攻撃に耐え、射精を我慢せよ～（072LABO様）	手コキ,オナニー,焦らし,強制/無理矢理,男性受け,洗脳	9016
【人類M男化計画】音声で手コキ 射精はゲームオーバー★マゾヒスト育成ゲーム「マゾヒストプランナー2・覚醒」～心菜さんに開放させられる～（072LABO様）	淫語,調教,言葉責め,強制/無理矢理,男性受け,洗脳	7579
【密着されて撫で回されて…】妖艶の湯～リラクゼーションオイルマッサージ～【サキュバス性感エステ】（M-STUDIO様）	マニアック/変態,ローション,言葉責め,羞恥/恥辱,男性受け,童貞	7225

- 相対的人気作（第三四分位数本～1999本）

タイトル	タグ構成	売上本数
【ハイレゾ】ダブルサキュバス～あなたのセイエキを耳元えっちでシボリつくしちゃう～【KU100】（テグラユウキ様）	癒し,淫語,ASMR,ロリ,手コキ,中出し	1949
【立体音響】Cure Maid-夕梨（ディーブルスト様）	癒し,バイノーラル/ダミヘ,ASMR,メイド,ラブラブ/あまあま,中出し	1913
あなたを壊す強制寸止めオナサポ～私のオモチャにしてあげる♪～（アルファートリル様）	淫語,オナニー,言葉責め,焦らし,逆レイプ,男性受け	1913

getBoxPlot_by_year.py

上のスクリプトは全期間の箱ひげ図を得るものなので推移がわかりませんでしょた. このスクリプトは年毎の売上を箱ひげ図で表します.



年毎の推移はfig/Boxplot_by_year.pngに, 各年の売上はBoxplot_[4桁数字].pngに生成されます.

上の図は砂糖様の過去5年間の活動の推移です. 2016年の箱が潰れているのは単純に出演数が少ないことが理由です. 2017年から2020年のいずれにおいても, 平均値が中央値を上回っていることが読み取れます.

この図はあくまでもその年に発売された作品の売上を表したもので, 年内売上を意味するものではありません. また, このデータは2020年10月に取得されたものなので2020年のデータは不完全である点に留意ください. 従って, 販売本数に関しては最近のデータは古いデータと比較して不利であると言えます.

※これより下, 取り扱うデータの性質上アダルトな内容が含まれます.

getTagData.py

このスクリプトはタグデータを集計し, 出現数, 出現率, 出現数/全タグの総和を算出します.

下図では出現数>=25を閾値として抽出します. なお, このデータでは『バイノーラル/ダミヘ, ASMR, 萌え』の3タグは性癖を示さず, 面白くないので除外しています.

タグ名	出現数	出現率	総数に占める出現数の割合
男性受け	37	26.24%	5.32%
淫語	32	22.70%	4.60%
ラブラブ/あまあま	29	20.57%	4.17%
手コキ	28	19.86%	4.02%
言葉責め	26	18.44%	3.74%
中出し	26	18.44%	3.74%

このデータはdata/getTagData.csvに生成されます.

『手コキ, ラブラブ/あまあま, 中出し』あたりはメジャーなタグと言えるでしょう. 特徴としては, 『男性受け, 淫語, 言葉責め』あたりのタグが多いように見えます. 一般論としてR18同人音声は, 視覚要素のある他の媒体と比較して男性受けが多いとされます. しかしながら, それを差し引いても特に男性受けタグの出現率が25%を超える方は珍しいという印象です. (製作者が調べた限り最も高かった方はA.K.様=30.88%なので, 砂糖様が群を抜いて高いという訳ではないです.)

ちなみにタグ『ロリ』と『お姉さん』は, 共に出現数16, 出現率11%の同率でした. これは, 砂糖様の演技の幅広さを客観的に示すことができるデータだと考えます.

これだけだと面白くないので, 売上ごとに使用タグの内訳に差があるか確認します. 今回は上位50%と下位50%に分割し, それぞれ示します. なお, 上記のデータとは取得時期が異なるため, 数が合わない点に留意ください.

- 売上本数上位50%のタグ構成

タグ名	出現数	出現率	総数に占める出現数の割合
男性受け	23	34.85%	6.32%
淫語	21	31.82%	5.77%
言葉責め	17	25.76%	4.67%
中出し	17	25.76%	4.67%
手コキ	15	22.73%	4.12%

- 売上本数下位50%のタグ構成

タグ名	出現数	出現率	総数に占める出現数の割合
ラブラブ/あまあま	19	29.23%	5.37%
手コキ	14	21.54%	3.95%
男性受け	14	21.54%	3.95%
中出し	12	18.46%	3.39%
お姉さん	12	18.46%	3.39%
耳舐め	12	18.46%	3.39%

結構傾向が別れたように見えます. 見た感じだと売上本数上位はM向けが多そうな印象があります. 一方で『手コキ, 中出し』あたりの一般的なタグは上位下位共に広い分布が見られました.

> 製作者の一言:
>
> 売上に関する話題はセンシティブなので, かなり慎重に言葉を選んだつもりです.
>
> あとこのデータは将来的に色々使えそうに思えます. そのうち他の方との比較や特徴の考察なんかを詳細に詰められたらいいなぁ, とか.

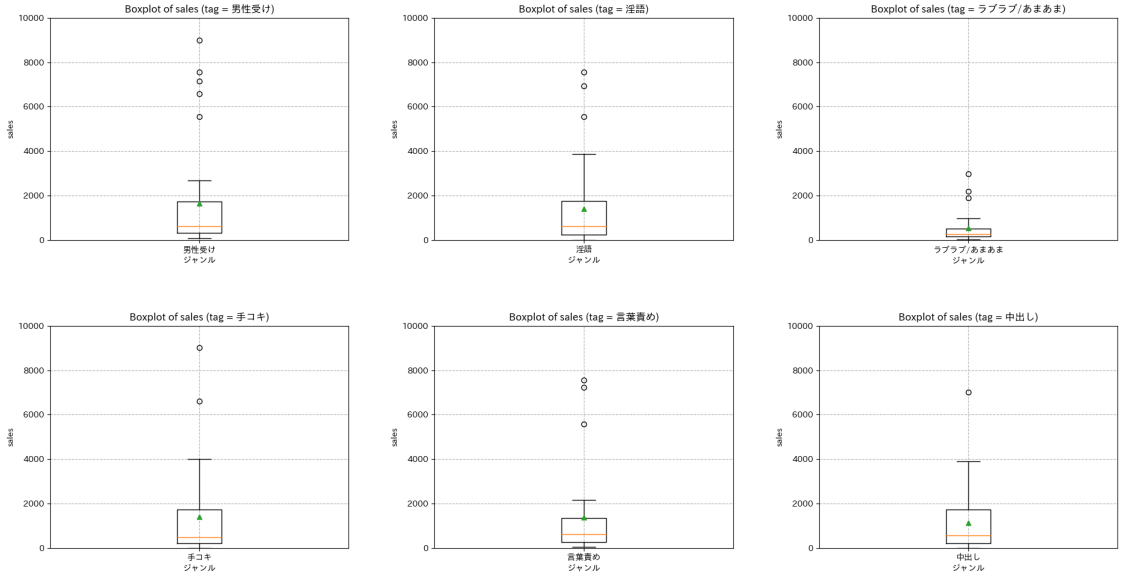
getTagBoxplot.py

このスクリプトは, タグごとの売上を箱ひげ図として出力します. 図を得たいタグはパラメタで指定します.

画像はfig/getTagBoxplot.pngに生成されます.

```
# ここでタグを指定する
data = data.query('tag.str.contains("ここにタグ名を書く")', engine='python')
```

今回は出現数数上位6タグを抽出します. 縦軸は比較のために同じスケールを使用します.



淫語, 手コキ, 中出しの上ひげは全体のそれを大きく上回ったものの, ラブラブ/あまあまのそれは全体とあまり変わりませんでした. この解析結果は中々興味深いです. もちろんラブラブ/あまあまにも相対的ヒット作（定義はgetBoxplot.pyの項目で解説）は存在しているので, 解析結果だけを見て『このタグは売れない』と言い切るのは短絡的でしょう.

getCircleCastData.py

このスクリプトは, 生データからサークルと出演声優のデータを抽出しランキングします.

まず, 砂糖様とかつて共演した声優様のデータを, 閾値>=3で以下に示します.

声優名	出現数	出現率
篠守ゆきこ	7	5%
大山チロル	6	5%
御崎ひより	4	3%
涼花みなせ	3	2%
逢坂成美	3	2%
秋山はるる	3	2%

このデータはdata/getCastData.csvに生成されます.

DLsite様利用者なら一度は耳にしたことのあるような重鎮ばかり. 各声優ごとに好きな作品は多数ありますが, それについて書くと今回の解説範囲を逸脱するので自重します.

今回はやりませんが, 声優ごとに共演データを収集したら面白いかもしれません.

続いて, 出演したサークル様のデータをトップ5まで示します.

サークル名	出現数	出現率
RIN world	13	10%
スタジオレイン	7	5%
072LABO	6	5%
テグラユウキ	5	4%
へーどねー	4	3%

このデータはdata/getCircleData.csvに生成されます.

こちらも錚々たる顔ぶれ. 各サークルごとに好きな作品は多数ありますが, それについて書くと今回の解説範囲を逸脱するので自重します.

その内, サークルごとの特色なんかを詳細に解析してみたいです.

注意:以下のスクリプトは独自研究を含みます. 話半分で見てください.

[試作]イメージ定量化の試み

この項目は, 何とかしてサークル様や声優様のイメージを量的に算出できないか?という試みです.

ここでは, イメージを定量化するためにイメージ係数という概念を導入します. イメージ係数とは, 解析対象のある作品Aが持つタグについて, 各タグの総数に占める出現数の割合(getTagData.py)の平均値を指します.

$$A = \left\{ t_1, t_2, \ldots, t_n \right\}$$

$$\text{イメージ係数}_A = \frac{\sum_{i=1}^{n} t_i \text{のタグ総数に占める出現数の割合}}{n}$$

この指標は,『解析対象が頻繁に使用するタグの用いられている作品は, 解析対象に対し購入者が抱くイメージに近いのではないか』という仮説に基づいて作成した指標です. なお, この指標は絶対的な尺度ではなく, 同一の条件で抽出されたデータ中で相対的に比較することを目的に作られています. つまり, **値が高い作品はこのデータ群において一般的な作風, 値が低い作品は珍しい作風**, のような見方ができます. 高いから良い, 低いから悪いという指標ではありません. (あくまでも作風の普遍性だけを見たいので, 売上による重み付けは行っていません.)

イメージ係数の最も高かった5作と最も低かった5作を抽出し, 以下に示します.

- Highest

作品名	タグ構成	ImageRate
征服少女と恋の宇宙戦争（amoroso様）	手コキ,パイズリ,男性受け	0.037
M for Manipulation スカーレット･ウィッチズ･ローズ編（夢想界様）	逆転無し,淫語,手コキ,足コキ,中出し,男性受け	0.034
甘えさせてくれるアンドロイドママ（インゴヒゴ様）	癒し,淫語,ラブラブ/あまあま,手コキ,中出し,ごっくん/食ザー	0.033
あなたを壊す強制寸止めオナサポ～私のオモチャにしてあげる♪～（アルファートリル様）	淫語,オナニー,言葉責め,焦らし,逆レイプ,男性受け	0.033
後輩女子のおち○ぽ奴隷へと成り下がる音声（アルファートリル様）	逆転無し,淫語,オナニー,言葉責め,逆レイプ,男性受け	0.033

イメージ係数最上位作品の傾向としては，女性が男性に対して優位に立つ展開の作品が多く見られるように感じました．

- Lowest

作品名	タグ構成	ImageRate
Story.3 傾国の白雪姫･中編 白雪親愛群像舞台（冷凍庫/freezer様）	健全,感動,シリーズもの,シリアス,金髪	0.003
彼女を脱がせて辱め～見られて感じる露出デート（ナンネット様）	女性視点,バイノーラル/ダミヘ,ASMR,色仕掛け,浮気,露出,羞恥/恥辱	0.004
変身ヒロイン悪堕ち報告書 総集編1（ドダメ屋さん様）	催眠,悪堕ち,洗脳,変身ヒロイン,けもの/獣化,フタナリ	0.006
黒い仔山羊の鳴く夜に～少女は嗤い、あなたを“マスター”と呼んだ～（思叫堂～ロア～様）	少女,人外娘/モンスター娘,年下攻め,ホラー,ロングヘア	0.006
【おっぱいず★】おっぱいは世界を救うのだ☆彡(私のエッチな歌で抜けぇぇー!シリーズ)（紙芝居屋さん十八軒目様）	ロリ,女教師,お嬢様,芸能人/アイドル/モデル,学校/学園,ツンデレ	0.007

イメージ係数が低い作品の傾向としては，人外，健全，女性視点や歌モノなどの比較的珍しい作風の作品が多く見られるように感じました．

以上のデータはdata/getImageRate.csvに生成されます．

イメージ係数と名付けたは良いものの，これをイメージだと言い切るのは少し論理の飛躍があるかもしれません．（ただし，高い=『解析対象にとってよく見られるタグ構成』，低い=『解析対象にとって珍しいタグ構成』ぐらいの事は言えます．）

この指標の特徴として，『上位タグを使用し，かつタグ使用数が少ない作品』が高く，『下位タグを使用し，かつタグ使用数が多い作品』が低くなる傾向が見られました．これは平均値利用の弊害と言えそうです．

また，サークル主様が設定していないタグやDLsite側で設定できない性癖にはどうしても対応できないという欠点があるため，指標の妥当性を過信するのは危険でしょう．

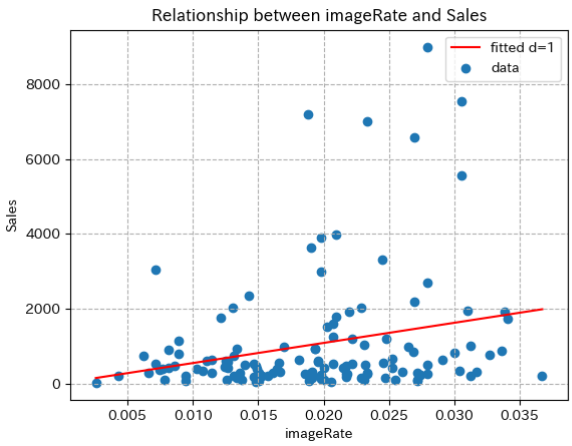
このままだと単なる自己満係数なので何かしらの方法で妥当性を確認したかったのですが…その方法が思いつきませんでした. 申し訳ないので自分を罵っておきます.

「ざこ❤ざぁこ❤クソザコ妥当性❤ガバガバ指標❤エビデンスよわよわ❤主観的な考察❤」

> 製作者の一言:
>
> 上ではボロカスに貶していますが, 単純な指標にも拘わらずそれっぽい結果が得られるので個人的には満足しています. この指標は以下のようなシナリオで用いると役立つと考えています. (と言うか製作者はこんな目的で活用しています.)
>
> - サークル様/声優様について, 普段の作風とは少し違う作品が聴きたい時, もしくは逆に出現頻度の高い作風の作品を聴きたい時
> - 新しいジャンルを, 好きなサークル様/声優様の作品で開拓したい時
> - 新しく知ったサークル様/声優様について, どのような傾向の作品が多いのかパッと知りたい時
>
> もちろん, あくまでこの指標は統計を用いた数字捏ね遊びの一環であり「これがこのサークル/声優らしい作品だ!」的な考えを押し付ける意図はありません.

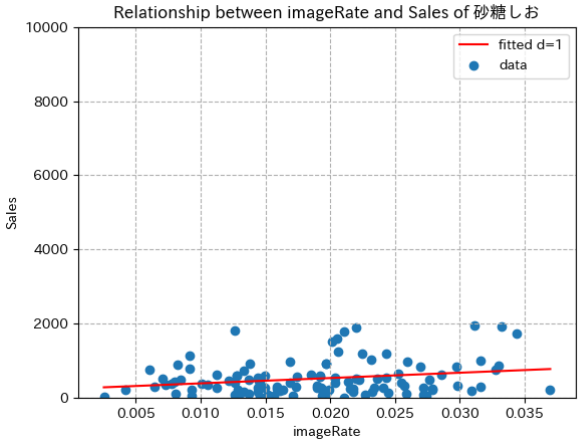
ついでなのでイメージ係数と売上の関係を調べてみましょう

という訳で, 先程得られたイメージ係数と売上をプロットし, 最小二乗法で近似した結果を以下に示します.



傾きと切片はそれぞれ[5.38485355e+04 7.51305653e+00], 相関係数は0.246でした. 傾きはかなり強いように見えますが, 相関係数を見る限り大した相関はなさそうです. つまり, イメージ係数は売上に大きく影響しないということです.

続いて, 外れ値(売上2000本以上の相対的ヒット作)を排除しプロットし直したものを以下に示します.

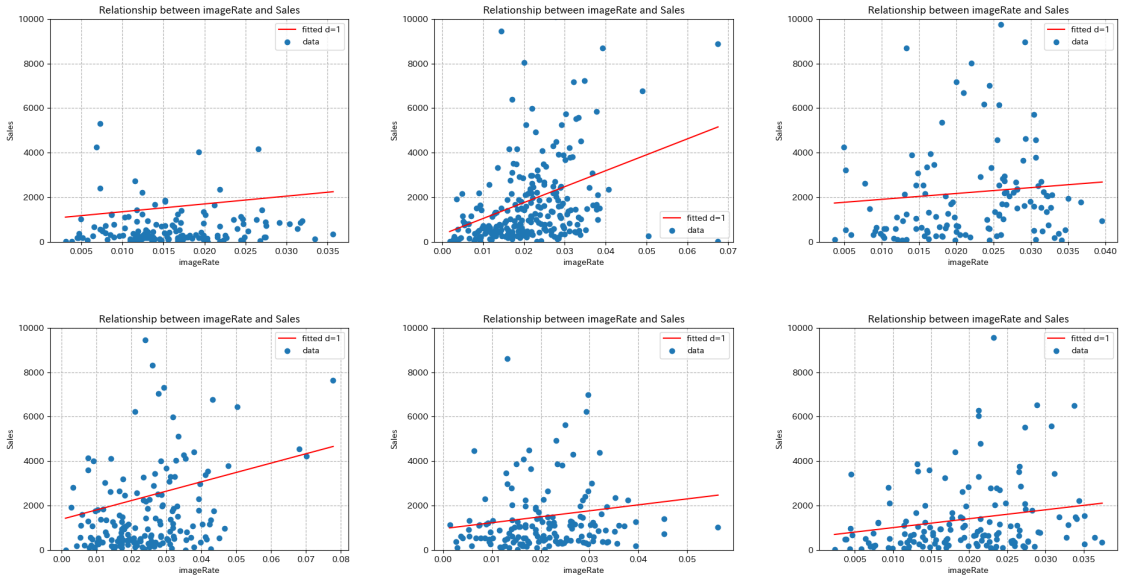


傾きと切片はそれぞれ[14294.67810006　251.66399918]と，かなり緩やかになります．相関係数は0.231と，殆ど変化ありません．

最小二乗法は外れ値に弱いとされます．つまり，相対的ヒット作が傾きを大きく変化させているということです．傾きが変化するのは想定の範囲内ですが，ここまで露骨に変化するとは思いませんでした．相対的ヒット作が正の傾きを強めているということは，相対的ヒット作の使用しているタグはイメージ係数の高いものが多いと言い換えることが可能です．もう少しわかり易い表現をするのであれば，『相対的ヒット作はメジャー性癖が多く，傾向的に保守的である』ということです．

相対的ヒット作を飛ばしているサークル様の傾向を見ると，所謂大手サークルが多い印象があります．中小サークルに話を限定すると，『マイナー性癖はそもそも需要が少ないので売れず，メジャー性癖では大手サークルに踏み潰されるので売れない』という，飽和した市場によくある現象が見られます．（もっとも，この議論では大手サークルと中小サークルの定義が曖昧ですが．）

ついでに，他の声優様6名で試しても，同様に正方向の傾きを持った近似式が得られました．相関係数は最大でも0.284だったので，ぶっちゃけ相関はほぼないと言って差し支えないでしょう．なお，ここではお名前は伏せさせて頂きます．



> 製作者の一言：
>
> 非常に個人的な話をすると，私は作品を買う時基本的にシチュで選んでいます．なので，相関係数が低いという結果は少し嬉しかったりします．

[試作]タグ共起関係の可視化

絞り込んだ作品から性癖同士の共起ネットワークを作成するスクリプトについての解説です．

まずは『getCo-OccNet.py』を実行して重み付きデータを得ます．実行結果の一部を以下に示します．このプログラムは，それぞれの行について生データを参照しながら共起関係を算出していくためけっこう重たいです．もしくは製作者の実装がまずくて必要以上に重くなっているかもしれません．

（当方環境では砂糖しお様で10000行/800sec.程度，最多出演を誇るY.H.様では20000行/5000sec.程度の実行時間でした．）

A	B	weight
男性受け	言葉責め	13
男性受け	手コキ	12
男性受け	淫語	10
⋮	⋮	⋮
足コキ	外国人	0
足コキ	獣耳	0
ツンデレ	触手	0

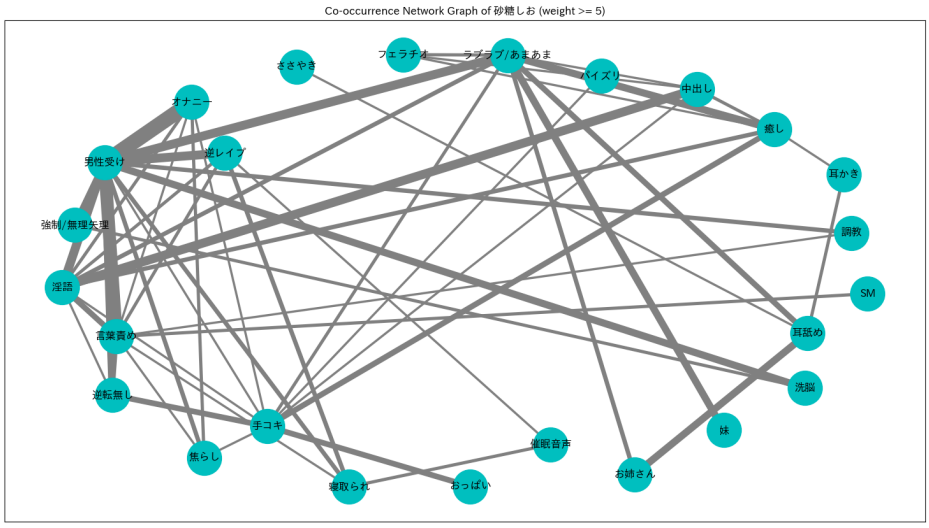
データはdata/co_occ.csvに生成されます.

このデータは共起関係とその頻度（登場回数）を示します. つまり例えば一行目からは,『男性受け』と『言葉責め』のタグが同時に登場する砂糖様出演作が累計13個あることが読み取れます.

結果としては, やはり『getTagData.py』の項目で議論したようなタグが上位に出ます.（この結果から性癖同士の相性を論じることはしません. なぜなら, そのような議論は声優個人のデータよりも, 全音声作品のデータを用いて行ったほうが妥当なものが行えるからです.）

今回は重み付けに単純な頻度を用いていますが, 適宜何かしらの係数（例えば, Jaccard係数は少しの改造で容易に実装できます.）を用いてもよいでしょう. 最終的にグラフに落とし込む際には, 流石に10000個の共起関係を全て描画するわけにも行かないので, ある程度の閾値を設けて絞り込みを行うべきです.

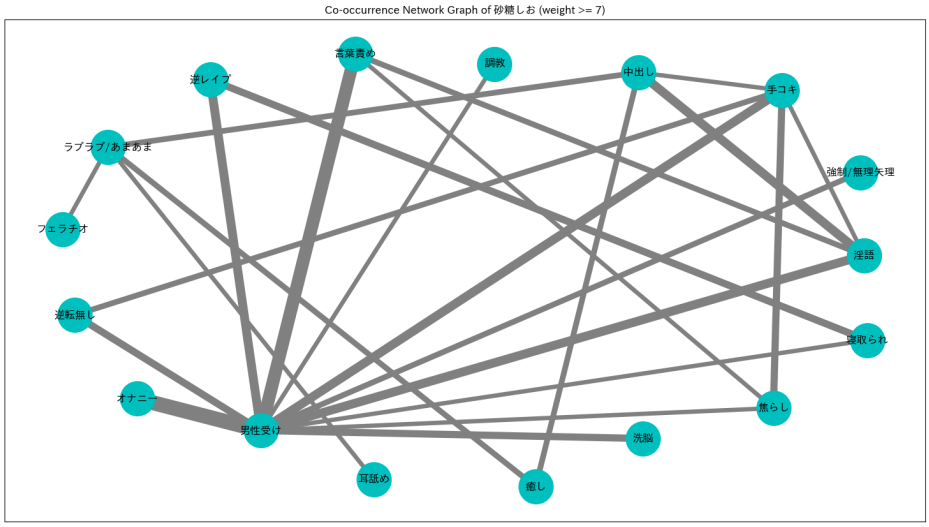
weightの閾値>=5で『getCo-OccNetGraph.py』を実行すると, 以下の図が生成されます.



先程のデータが図表として整理され, 視覚的な理解が容易になりました. 強い共起関係ほど太い線で表されます. 図の中では主に『男性受け, 手コキ, ラブラブ/あまあま, 淫語』あたりのタグが多くのタグと共起関係を持っていることが読み取れます.

あくまでも見た感じの感想ですが,『ラブラブ/あまあま, 癒し, 中出し, 耳舐め, お姉さん』といった比較的一般的なジャンルグループと,『男性受け, 言葉責め, 逆転無し, 強制/無理矢理』といった比較的Mな視聴者向けのジャンルグループに大別されているような印象を持ちました.

上の図は少し見づらいので，閾値>=7に上げて再生成したものを下に示します．



閾値を上げると，多くのタグが『男性受け』タグに対する共起関係を持っていることが読みれます．

上に来るのは一般的なタグだろうと思っていたのですが，意外にも特徴が色濃く反映されていて面白いという感想でした．特に強い共起関係を持ったタグについては，ファンであれば『あの作品のことか』なんて思う物があるのではないでしょうか？

この図は単純に眺めていて面白いので，ご自分の好きなサークル様/声優様で作ってみたらいいと思います．

> 製作者の一言：
>
> ...もう少しこの辺の知識があれば「このタグとこのタグは共起無いけど相性良さそうだぜへへン」とか「このタグを入れることでメッチャ売れるぜフフン」とか「こういう傾向だとこの人の魅力を最大限引き出せるぜハハン」みたいな格好いい考察ができるんでしょうか？
>
> この辺の技術って自然言語処理などでは頻繁に出てくるらしいんですが，ぶっちゃけ門外漢なのであまり詳しくないです．つよつよな方がいらっしゃいましたら，どうかお手柔らかにお願いします．

[試作]タイトルへの形態素解析

続いて，砂糖様出演作に対して形態素解析を行い，出演作の傾向を探ってみましょう．

今回は最も主流の日本語形態素解析ツールであるMeCabを用います．対応しきれない淫語については適宜ユーザー辞書を作成します．まず具体的なタイトルに対して形態素解析を行った結果を示します．この例では砂糖様個人サークルであるしゅがぁそると様の作品『【KU100】姉サキュ～弟くんへの独占欲が強すぎてサキュバスになった姉に管理されちゃう～』を事例として用います．

```
> mecab
【KU100】姉サキュ～弟くんへの独占欲が強すぎてサキュバスになった姉に管理されちゃう～
【    記号,括弧開,*,*,*,*,【,【,【
KU   名詞,一般,*,*,*,*,*
100 名詞,数,*,*,*,*,*
】    記号,括弧閉,*,*,*,*,】,】,】
姉    名詞,一般,*,*,*,*,姉,アネ,アネ
```

```
サキュ  名詞,一般,*,*,*,*,*
～　　名詞,サ変接続,*,*,*,*,*
弟　　名詞,一般,*,*,*,*,弟,オトウト,オトート
くん　名詞,接尾,人名,*,*,*,くん,クン,クン
へ　　助詞,格助詞,一般,*,*,*,へ,へ,エ
の　　助詞,連体化,*,*,*,*,の,ノ,ノ
独占　名詞,サ変接続,*,*,*,*,独占,ドクセン,ドクセン
欲　　名詞,一般,*,*,*,*,欲,ヨク,ヨク
が　　助詞,格助詞,一般,*,*,*,が,ガ,ガ
強　　形容詞,自立,*,*,形容詞・アウオ段,ガル接続,強い,ツヨ,ツヨ
すぎ　動詞,非自立,*,*,一段,連用形,すぎる,スギ,スギ
て　　助詞,接続助詞,*,*,*,*,て,テ,テ
サキュバス　　名詞,一般,*,*,*,*,*
に　　助詞,格助詞,一般,*,*,*,に,ニ,ニ
なっ　動詞,自立,*,*,五段・ラ行,連用タ接続,なる,ナッ,ナッ
た　　助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
姉　　名詞,一般,*,*,*,*,姉,アネ,アネ
に　　助詞,格助詞,一般,*,*,*,に,ニ,ニ
管理　名詞,サ変接続,*,*,*,*,管理,カンリ,カンリ
さ　　動詞,自立,*,*,サ変・スル,未然レル接続,する,サ,サ
れ　　動詞,接尾,*,*,一段,連用形,れる,レ,レ
ちゃう  動詞,非自立,*,*,五段・ワ行促音便,基本形,ちゃう,チャウ,チャウ
～　　名詞,サ変接続,*,*,*,*,*
EOS
```

特に解説することもありませんが, 品詞の分類, 読み方, 活用などが返ってきます.

では以下に,『getTitleNoun.py』を用いて調べた名詞出現率トップ10を示します. せっかくなので他の方と比べてみましょう. なお, ここではお名前は伏せさせて頂きます. 当ててみてください. (A.K.様に関する統計は改名前, 改名後の和集合です.)

No.	砂糖しお/単語	砂糖しお/出現率	A.K./単語	A.K./出現率	M.S./単語	M.S./出現率
1	音声	18.0	バイノーラル	14.4	バイノーラル	26.2
2	射精	14.1	催眠	13.0	耳	18.0
3	マゾ	13.3	音声	13.0	音声	12.2
4	バイノーラル	12.5	姉	10.6	KU100	12.2
5	さん	12.5	さん	10.2	メイド	8.7
6	奴隷	11.7	耳	9.5	JK	7.6
7	オナニー	10.9	様	8.5	さん	7.6
8	ちゃん	10.2	調教	8.1	射精	7.0
9	編	8.6	サキュバス	7.7	姉	7.0
10	姉	7.8	2	7.4	ちゃん	7.0

No.	R.I./単語	R.I./出現率	K.A./単語	K.A./出現率	Y.H./単語	Y.H./出現率
1	ちゃん	12.6	バイノーラル	16.2	バイノーラル	20.3
2	後輩	12.0	JK	14.3	耳	12.5
3	JK	10.8	KU100	12.1	姉	11.9
4	音声	10.8	えっち	10.2	ちゃん	11.0
5	KU100	9.0	音声	9.5	さん	10.6
6	バイノーラル	9.0	耳	8.3	KU100	10.6
7	耳	9.0	フォーリー	7.6	JK	9.1
8	メイド	7.8	様	7.6	音声	9.1
9	彼女	7.8	ギャル	7.0	メイド	7.3
10	さん	7.8	さん	7.0	ASMR	6.7

主観的な感想になりますが，各人の出演作の特徴を色濃く反映しているように見えます．（例えば『A.K.様=催眠・調教』，『R.I.様=後輩・彼女』，『K.A.様=ギャル・フォーリー』なんてそのまんまです．）このデータはかなりいいおもちゃになりそうな気がします．タグの出現頻度と組み合わせたら，より高精度で特徴算出ができるかもしれない．

また，非常に馬鹿馬鹿しいですがタイトル自動生成とか作ったら笑えるかもしれない．アルゴリズム的には，『（人物）の（プレイ傾向）な（プレイ内容）の音声』みたいな単純なものであればすぐ作れそう．

> 製作者の一言：
>
> アダルトコンテンツは淫語が多いという性質上，未知語に対するロバスト性が高いKyTea（って情報を何かの論文で読んだ）を用いても面白いかもしれないです．どうでもいいけど何で日本語のNLPツールって変な名前のが多いんだろうね．（MeCab然りCaboCha然り）

ついでなので動詞と形状詞のランキングも掲載します．

No.	動詞/単語	動詞/出現率	形状詞/単語	形状詞/出現率
1	さ	17.6	エッチ	5.3
2	し	12.2	えっち	5.3
3	する	7.6	的	3.1
4	くれる	6.1	オーバー	3.1
5	堕ち	6.1	あまあま	2.3
6	寝取ら	5.3	イキ	2.3
7	舐め	5.3	大量	1.5
8	なっ	3.8	不貞	1.5
9	甘やかし	3.1	クール	1.5
10	果てる	2.3	淫乱	1.5

いずれこれを使って何かするかもしれません.

[試作]タイトルへの係り受け解析

折角なので, 係り受け解析ツールCaboChaを用いてタイトルに対する構文解析を行います. 事例は同上.

```
【KU100】姉サキュ～弟くんへの独占欲が強すぎてサキュバスになった姉に管理されちゃう～
【KU100】姉サキュ～弟くんへの-D
                        独占欲が-D
                          強すぎて---D
                        サキュバスに-D
                                  なった-D
                                      姉に-D
                          管理されちゃう～
EOS
```

『～』や『【】』の解析がいまいち上手く行っていませんが, とりあえず文書構造は分かりました.

解析結果を見ると, この文章は動作の対象が省略されている点に特徴づけられます. (対象=主人公≒消費者であることが暗黙の約束となっている.)また, この文章は『各文節が直後の文節を修飾する文書構造』になっており, 頭から読み進めたときに内容の理解が容易です. そのため, このような文書構造はプロットを消費者に短時間で理解させることに有利であり, 特にライトノベルのタイトルでも頻繁に用いられています.

今はまだNLP勉強中なので深いことができませんが, いずれこれを使って何かするかもしれません. 同人音声のタイトル解析向けに調整しても面白いかもしれませんし, 売上上位作品のタイトル構文にどのような傾向が見られるかを調べてみても面白いかもしれません.

> 製作者の一言:
>
> あんま関係無いんだけど,『寝取られ』がちゃんと正しく分解されるのに笑ってしまった.
>
> ```
> 僕は-------D
> 最愛の-D    |
> ```

```
          妻を---D
          部下に-D
      寝取られた。
EOS

* 0 4D 0/1 3.900490
僕    名詞,代名詞,一般,*,*,*,僕,ボク,ボク 0
は    助詞,係助詞,*,*,*,*,は,ハ,ワ     0
* 1 2D 0/1 1.621914
最愛   名詞,一般,*,*,*,*,最愛,サイアイ,サイアイ   0
の    助詞,連体化,*,*,*,*,の,ノ,ノ     0
* 2 4D 0/1 5.514941
妻    名詞,一般,*,*,*,*,妻,ツマ,ツマ    0
を    助詞,格助詞,一般,*,*,*,を,ヲ,ヲ    0
* 3 4D 0/1 0.000000
部下   名詞,一般,*,*,*,*,部下,ブカ,ブカ   0
に    助詞,格助詞,一般,*,*,*,に,ニ,ニ    0
* 4 -1D 0/2 0.000000
寝取ら  動詞,自立,*,*,五段・ラ行,未然形,寝取る,ネトラ,ネトラ 0
れ    動詞,接尾,*,*,一段,連用形,れる,レ,レ 0
た    助動詞,*,*,*,特殊・タ,基本形,た,タ,タ     0
。    記号,句点,*,*,*,*,。,。,。  0
EOS
```

[試作]単語が売上に与える影響

面白くなってきたのでもう少し深く解析してみます. ずっと気になっていたのですが, 【バイノーラル】【KU100】【ハイレゾ】みたいな単語をタイトルにつけて謳い文句にしている作品は多いです. 実際のところこれは売上に影響しているのでしょうか?『wordsEffect.py』を用いて調べてみます.

という訳で調査します. 帰無仮説と対立仮説はそれぞれ以下の通り.

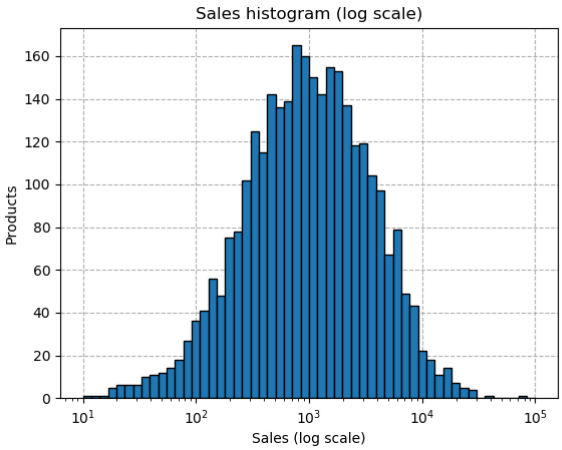
- H_0:【バイノーラル】【KU100】【ハイレゾ】が付いている作品の売上は, 付いていない作品と差がない.
- H_1:【バイノーラル】【KU100】【ハイレゾ】が付いている作品の売上は, 付いていない作品と差がある.

今回は本マニュアルで登場した声優様+2019年の出演数上位の声優様計16名の出演作に絞ります. なお, ここではお名前は伏せさせて頂きます.

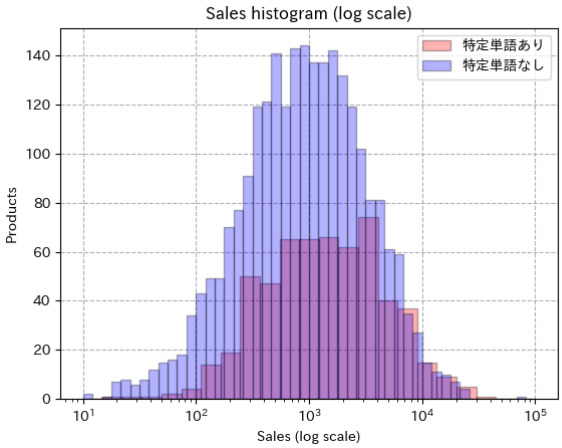
では得られたデータの要約統計量を示します.

```
             price          sales
count  3040.000000    3040.000000
mean   1042.305592    2248.235197
std     374.317772    6974.322499
min       0.000000       0.000000
25%     880.000000     405.750000
50%    1100.000000     982.000000
75%    1210.000000    2356.500000
max    4400.000000  249125.000000
```

全体をヒストグラムにするとこんな感じ. 思ったより綺麗になった. 相変わらず10^3程度が売上の中央値です.



続いて，タイトルに【バイノーラル】【KU100】【ハイレゾ】と付いている作品と付いていない作品をそれぞれ抽出し，ヒストグラムを以下に示します．使用したスクリプトは/script/wordsEffect.pyにあります．



有意水準0.05でMann-WhitneyのU検定を行った結果を以下に示します．

```
MannwhitneyuResult(statistic=848513.5, pvalue=1.7699014834094937e-12)
```

$p < 0.05$となり，よって帰無仮説は棄却されます．即ち，『【バイノーラル】【KU100】【ハイレゾ】が付いている作品の売上は，付いていない作品と差がある』と言えます．これは，前記のような単語をタイトルに付けていないサークル様にとっては，少々不本意な結論になるかもしれません．

> 製作者の一言：
>
> データ総数は3040件なので，**『一期一会・一日一発』で8.3年掛かる計算です．**
>
> 売上平均は2248本でなので，『購入者全員が1作品で1回致し，1回あたり精子3億匹』と仮定するのであれば，ゴミ箱に捨てられた精子は2.05e+15匹です．2000兆匹…ちょっと天文学的な数字過ぎて実感が湧きません．
>
> 『1回3ml』と仮定するならば2万リットル．4Kリットルタンクローリー5台を満載にする量です．人間が一生に飲む水の量が4万リットルと言われるので，飲料水として用いるのであれば人生の半分を賄えます．
>
> ちなみに，2万リットルのジェット燃料はB747-400を95分間飛ばすことが可能です．東京から那覇まで行けます．

手法の限界

- 声優様について

  今回使用したスクリプトでは出演者タグが付いていない製品をカウントしていないので，数え漏れが発生する場合があります．

  また，モブ役としての出演も主人公としての出演も統計上では同等に扱われていることに留意すべきです．

  出演作情報から得られるのはあくまでも発売日です．実際の収録時期との間にはズレがあります．

- サークル様について

  サークル様につきましては，どうしても声優様と比較してデータ量が少なくなりがちな傾向があります．

  今回はサークル様についての調査を行っていないので，あまり書けることがないです．

- タグ情報について

  サークル様が設定していないタグやDLsite様に実装されていないタグには対応できないという手法の限界があります．

  製作者もマイナー性癖の検索には苦労しているので，この辺り（特にNLPとかの領域）に詳しい方いましたら教えてください．

今後の展望

今後やりたいこととか感想とか

類似度算出

サークル様/声優様同士の類似度を出す的なやつです．

ある程度の特性はタグやタイトルの分析で把握できるので，アルゴリズム的にはそこまで難しくはないでしょう．ただし全部集めるとなると労力が半端なくマゾいので，「今後やりたい」程度に留めておきます．

分類の詳細化

もうちょっと分類を詳細化したいというやつです．

DLsite様のタグは痒い所に手が届かない部分があることは利用者の皆様も知ることでしょう．（別に悪口を言っているわけではありません．それだけ多様な性癖向けの作品が投稿されている証左であり，とても喜ばしいことだと思います．）

ただ現状ではマイナー性癖の作品を探すのに結構苦労するので，何かしらの方法で改善を試みたいです．

robots.txtで各製品ページやレビューに対するアクセスが禁止されているため，実装方法は決めかねています．

性癖偏差値の導出

これは分析と言うよりはネタに走っています．

勿論元ネタは皆様ご存知あのコピペです．アレをDLsite様の統計に基づいて作ってみよう的なやつです．

要するに一般的な性癖は低く，特殊性癖ほど高くなるようにすれば良いので，各タグごとに出現頻度を集計し，何かしらの方法でこねくり回して何かしらの方法でスコア化すれば可能なんじゃないでしょうか．（100-出現率とか，出現率の逆数を取るとか．）

製作者はクリエイターではないので，実際のクリエイター様がどんな機能を求めているのかよく分かりません．

こんな機能があったら便利だと思うよってアイデアをお持ちの方は，是非GitHubのIssue等で教えてください．

製作者の感想

このツールはもともと個人用途で作ったやつなのでソースやアルゴリズムが汚いかもしれないです．

さて，少なくとも製作者が属している世界では，アダルトコンテンツを扱った研究はフィルタリング目的が大半で，『完全100%エロ目的』って研究は見たことないです．まぁまともな研究機関ならそんなものに予算降りるはずがないので，当たり前といえば当たり前の話です．しかしながら，これだけ歴史があって大きなエロ音声市場というのは特異であり，研究対象として非常に価値が高いと考えています．

あと，DLsite様の音声作品全体のデータを解析したり，購入履歴から利用者の性癖を解析するみたいなネタは先行研究としてありました．しかしながら，声優様やサークル様を解析対象として，それらの活動支援を目的としたツールというのは（製作者が探した限りでは）見つかりませんでした．是非サークル様・声優様に本ツールをご活用いただき，今後の活動の助けとして頂ければ幸いでございます．また，DLsiteユーザ様にはオカズ探し用のツールとしてご活用頂けますと幸いです．

ネット上の文化を対象とした統計解析はやっていて楽しいしネタとしても面白いので，今後も断続的に行っていきたいです．

謝辞

本マニュアルの解析結果は，砂糖しお様にご許可を頂いた上で掲載しております．

ご多忙中にも拘わらず，快いご対応を頂き深く感謝いたします．

砂糖しお様の今後の益々のご活躍と，音声作品業界の発展をお祈り申し上げます．

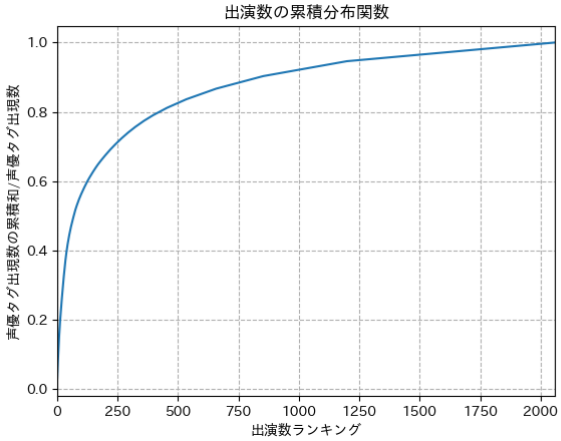
付録

どうせなので全体に対しても解析を行ってみたくなりました．2020年11月18日に取得した全作品のデータに対する解析を行った結果を付録として付けておきます．おまけの中身が知りたい方のみご参照ください．

	price	sales
count	24439.000000	24439.000000
mean	932.493146	840.414624
std	1004.677342	3956.606518
min	0.000000	0.000000
25%	550.000000	40.000000
50%	880.000000	184.000000
75%	1100.000000	639.000000
max	44000.000000	252370.000000

データ総数は24439，売上本数の中央値は184，平均は840です．このデータからも，平均と中央値には大きな剥離があることが見て取れます．

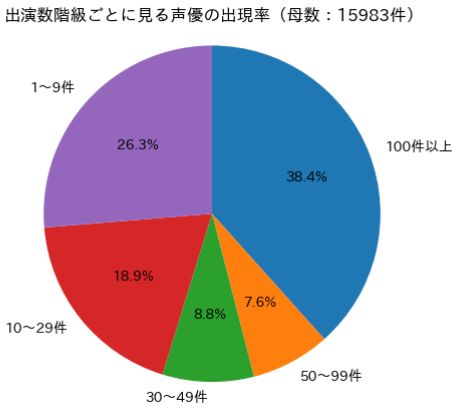
全期間での声優タグの登場数は15983，声優の総数は2059でした．まずは出演数のランキングをx軸として累積分布関数を示します．なお，別名義や改名を考慮していない点に留意ください．



うわぁ気持ちいいぐらい露骨に指数分布ってますね，これは．

『有名声優に多少は集中してるんだろうな～』的な予想はしていたのですが，ここまで露骨なグラフが出るのは衝撃でした．データから言えば，上位5名で10%，13名で20%，40名で40%，125名で60%，422名で80%の依頼が回っている計算になります．ちなみに，今回データを拝借した砂糖しお様は出演数ランキング第26位です．これは上位1.3%に相当します．

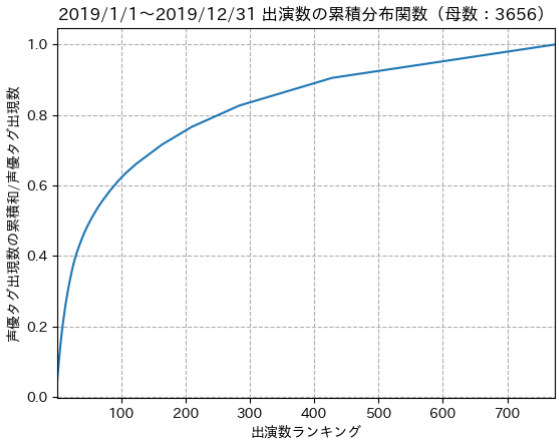
このグラフだとちょっと分かり辛いので，出演数ごとに階級を分け，勢力を円グラフに表します．



出演数100件超の声優36名（上位1.7%）が作品全体の38.4%を独占しているのに対し，出演数1～9件の声優1746名（下位84.8%）が26.3%を食い合っています．サークル様だけでなく，一見きらびやかに見える声優達もえげつない競争に晒されていることが理解できます．

ドラえもんのスネ夫役で有名な関智一氏の著書『声優に死す　後悔しない声優の目指し方』によれば，声優を志望する者は毎年約30000人，その中で日俳連のジュニアランクに登録される者は200人程度いるとのことです．（この時点で1%未満！）そして，この激しい競争は全年齢の世界だけでなくR-18の世界にも当てはまるようです．

もっと世知辛い話をしましょう．以下の累積分布関数は，同様のものを2019年のデータから作成したものです．



文字単価4円*2万5千文字=10万円をモデルケースと仮定するならば，依頼のみで年収250万円に到達するには年25本，400万円に到達するには年40本の出演が必要となります．**2019年のデータを見ると，年間出演数25本超は27名（上位3.5%），40本超に至ってはたったの16名（上位2.1%）でした．**（もちろんこの試算はあくまでもモデルケース*出演数のみを考慮しているので，実態とは剥離があります．例えば自サークル運営による収入は考慮していません．）

確かに，全年齢向けに活動している声優の数と比較すると『R-18で個人・同人を相手に活動している声優』はかなり少数派であり，市場自体は狭いと言えます．しかしながら，その中で『継続的に活動し，なおかつ専業で飯を食えている声優』になるためには，茨の道と形容するのも生温い程の競争を制する必要があります．

続いて，直近2年分のデータを以下に示します．2020年のデータは11月18日に取得されたものであり，不完全である点に留意ください．

```
2020年（11月18日時点）：
              price          sales
count   4063.000000    4063.000000
mean     951.355402     891.030273
std      852.583816    3429.803149
min        0.000000       0.000000
25%      550.000000      27.000000
50%      990.000000     144.000000
75%     1210.000000     622.000000
max    33000.000000  141980.000000

2019年：
              price          sales
count   3845.000000    3845.000000
mean     886.672562    1346.619766
std      783.291081    7362.458724
min        0.000000       0.000000
25%      550.000000      51.000000
50%      880.000000     232.000000
75%     1100.000000     947.000000
max    33000.000000  223071.000000
```

まず，作品数は11月18日時点で昨年比+5.7%と，相変わらずの成長を続けています．

平均売上本数は昨年比-33.8%，中央値は昨年比-37.9%と，いまいち伸び悩んでいるように見えます．製作者は『COVID-19の影響によって家にいる機会が増えたとを背景に音声作品の売上が増加する』と見込んでいたのですが，どうやら見当違いだった様です．販売期間の差は当然あると思いますが，それを差し引いても中央値-37.9%という値は異常に見えます．

2020年の販売価格の中央値は990円と，昨年と比較して110円の高値となりました．これは2010年以来10年ぶりの高値です．たった110円と思われますが，割合にすると+12.5%です．ユーザによってはかなり値上がりを体感した1年なのではないでしょうか？

タグ出現数トップ10を示します．

No.	タグ名	出現数	出現率
1	フェラチオ	5057	20.69
2	癒し	4464	18.27
3	淫語	4000	16.37
4	中出し	3766	15.41
5	ラブラブ/あまあま	3752	15.35
6	オナニー	3297	13.49
7	言葉責め	3171	12.98
8	手コキ	2741	11.22
9	男性受け	2443	10.00
10	耳かき	2202	9.01

全体的な傾向を見ると，思ったよりも様々な需要が有ることに気付かされます．ある程度大別するのであれば，『フェラチオ，中出し』あたりは比較的ノーマルなジャンルと言えるでしょう．『言葉責め，男性受け』のようなタグはマゾ向けです．男性が布団に寝転がって聴くという性質上，男性が受けに回るのはある程度必然です．『癒し，耳かき，ラブラブ/あまあま』のように，性的な欲求だけではなく，心の安らぎを求める需要も根強いようです．『淫語，オナニー，手コキ』の付いている作品は，所謂『抜き特化』の傾向が見られました．

名詞，動詞，形状詞の出現率ランキングは以下のとおりです．

No.	名詞/単語	名詞/出現率	動詞/単語	動詞/出現率	形状詞/出現数	形状詞/出現率
1	さん	8.01	し	8.18	エッチ	2.63
2	様	6.86	さ	6.58	えっち	2.39
3	バイノーラル	6.65	舐め	3.21	フリー	1.79
4	姉	6.46	する	2.71	的	1.17
5	音声	5.95	あげる	1.28	大好き	1.14
6	ちゃん	5.79	なっ	1.24	逆	0.97
7	耳かき	5.25	くれる	1.20	好き	0.90
8	2	5.00	み	0.83	あまあま	0.85
9	催眠	4.84	ま	0.83	リアル	0.78
10	射精	4.78	い	0.81	エロ	0.71

名詞に関してはみんなお姉ちゃん好きなんだなぁと思った. 俺もソーナノ.

動詞に関しては『舐め』は多分耳舐めのことかな?確かに多い.『あげる, くれる』みたいな受け身の動詞が多いのは音声作品特有かなとおもった

形状詞に関しては『大好き, 好き, あまあま』など, タグと比較するとかなり甘めの傾向に見えました.『エッチ, えっち, エロ』の違いはよくわかんない. にほんごむつかしいね.