連絡

論文紹介について

- 要点をまとめた簡単なレジュメを人数分用意してきてください。
- 授業で論文の全てのセクションを順番に「読む」必要はありません。本筋に関係しないと思うところは飛ばしてもかまいません。
- 各論文は事前にアクセス可能にします。

論文紹介について

- とくに次のような点に注目してまとめてください。論文の内容を批判的に読むこと、自分自身の研究に生かすことができるかを意識してください。
  - 調べたいトピックは何か？仮説は何か？
  - 使ったコーパスは何か？それは調べたいトピックに合っているか？
  - コーパスをどのように調べているか？手作業か、なにか特殊な技術、ソフトウェアを使っているか？
  - 調査されていないこと、ロジックの問題点、さらに発展させるべき点はないか？

正規表現

コーパス調査入門 (淺尾)

2015年5月20日

おしながき

- 正規表現とは何か
- 正規表現のいろいろ

正規表現

- 「高度な検索」を行う時に使う
  - 「数字で始まる行」を検索するとか
  - ある文字が含まれていない行を検索するとか
  - 前後を《》で囲まれた文字のみ検索するとか
- いろいろな場面で使える
  - テキストエディタ
  - コーパスのソフトウェア（コンコーダンサ）
  - プログラミング言語

チョムスキー階層

- 帰納的可算言語
- 文脈依存言語
- 文脈自由言語
- 正規言語
  - 正規言語を表現するのが正規表現

正規表現を書こう！

?

- 直前の文字があってもなくてもいい
- てい?る
  - 「てる」と「ている」の両方にマッチ

.

- 任意の文字にマッチ
- の.の
  - 「の村の」「の手の」「のんの」「の☆の」などにマッチ

+

- 直前の文字の繰り返し
- あ+
  - 「あ」「ああ」「あああああああ」などにマッチ

*

- 直前の文字の０回以上の繰り返し
- すごー*い
  - 「すごい」「すごーい」「すごーーーい」などにマッチ

行頭と行末

- ^ : 行頭
- $: 行末
  - 。$: 行末にある「。」にマッチ
  - ^The : 行頭にある The という文字列にマッチ

問題

- 「コンピュータ」と「コンピューター」に同時にマッチする正規表現は？
- ! で終わる行にマッチする正規表現は？
- 羊の発話（ba, baaaa, baaaaaaaaa など）にマッチする正規表現は？

[ ]

- 複数候補
- [こそあど]れ
  - 「これ」「それ」「あれ」「どれ」にマッチ

[ ]

- 範囲指定もできる
  - [0-9] : 数字ぜんぶ
  - [A-Za-z] : アルファベットぜんぶ
  - [ぁ-ん] : ひらがなぜんぶ
  - 漢字ぜんぶ…はちょっと難しい（文字コード依存）

[ ]

- [^] で否定もできる
  - [^a] : a 以外
  - [^、。！？] : 句読点など以外
  - [^0-9] : 数字以外

|

- 赤シャツ|山嵐
  - 「赤シャツ」「山嵐」の両方にマッチ
- (赤シャツ|山嵐)が
  - 「赤シャツが」「山嵐が」の両方にマッチ
  - 赤シャツが|山嵐が と書いても同じ

その他よく使うもの（主に英語向け）

- \n : 改行
- \t : タブ
- \s : スペース類（タブ、改行含む）
- \b : 単語境界（これ自体はゼロ文字）
- \w : アルファベット [a-zA-Z]
- \W : アルファベット以外 [^a-zA-Z]
- \d : 数字 [0-9]
- \D : 数字以外 [^0-9]

問題

- 「これ」「それ」「あれ」「どれ」、「この」「その」「あの」「どの」の全てにマッチする正規表現は？
- 行の初めの数字の連続にマッチする正規表現は？

問題

- *the* という単語にマッチする正規表現は？
- *as long as* など、単語 1 個を挟んで as が2つあるところを検索するには？