

エクテック カジュアル勉強会

Pythonで自然言語処理 & トピックモデルを学ぶ

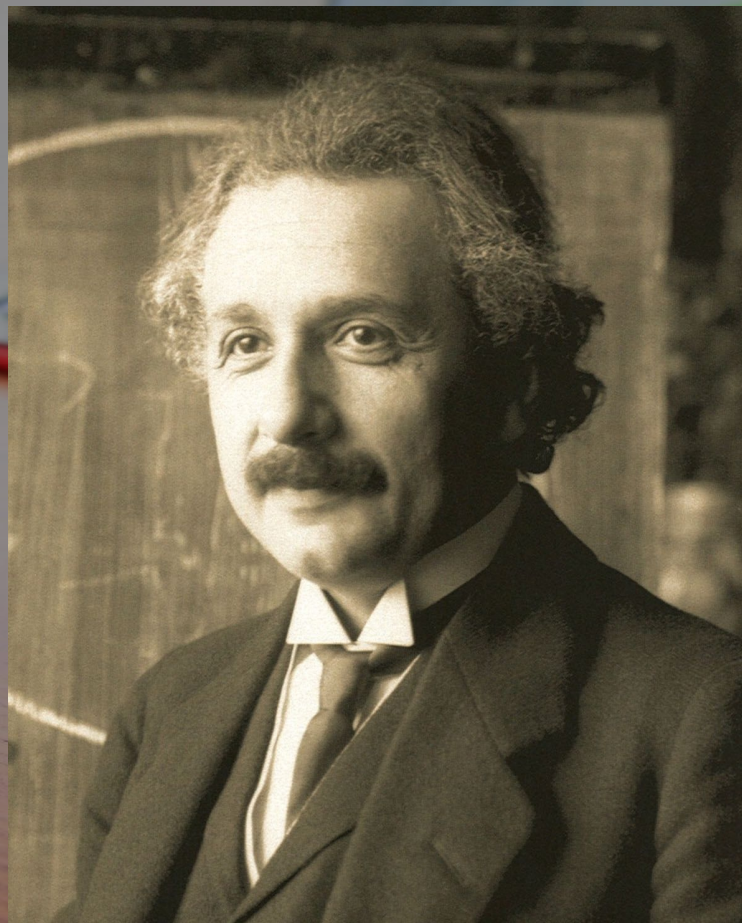
A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a red and white patterned pen, and an open book with French text. The scene is set on a wooden desk.

エクテック カジュアル勉強会 趣旨と概要

le jour à Paris dans les années 1860. Plutôt qu'
vestimentaires de la bourgeoisie, les créateurs
courant de l'esthétisme.

エクテック カジュアル勉強会 趣旨

限りある時間の中でも
様々な見識や知見を吸収できないか？
※興味や意欲、好奇心への刺激一環



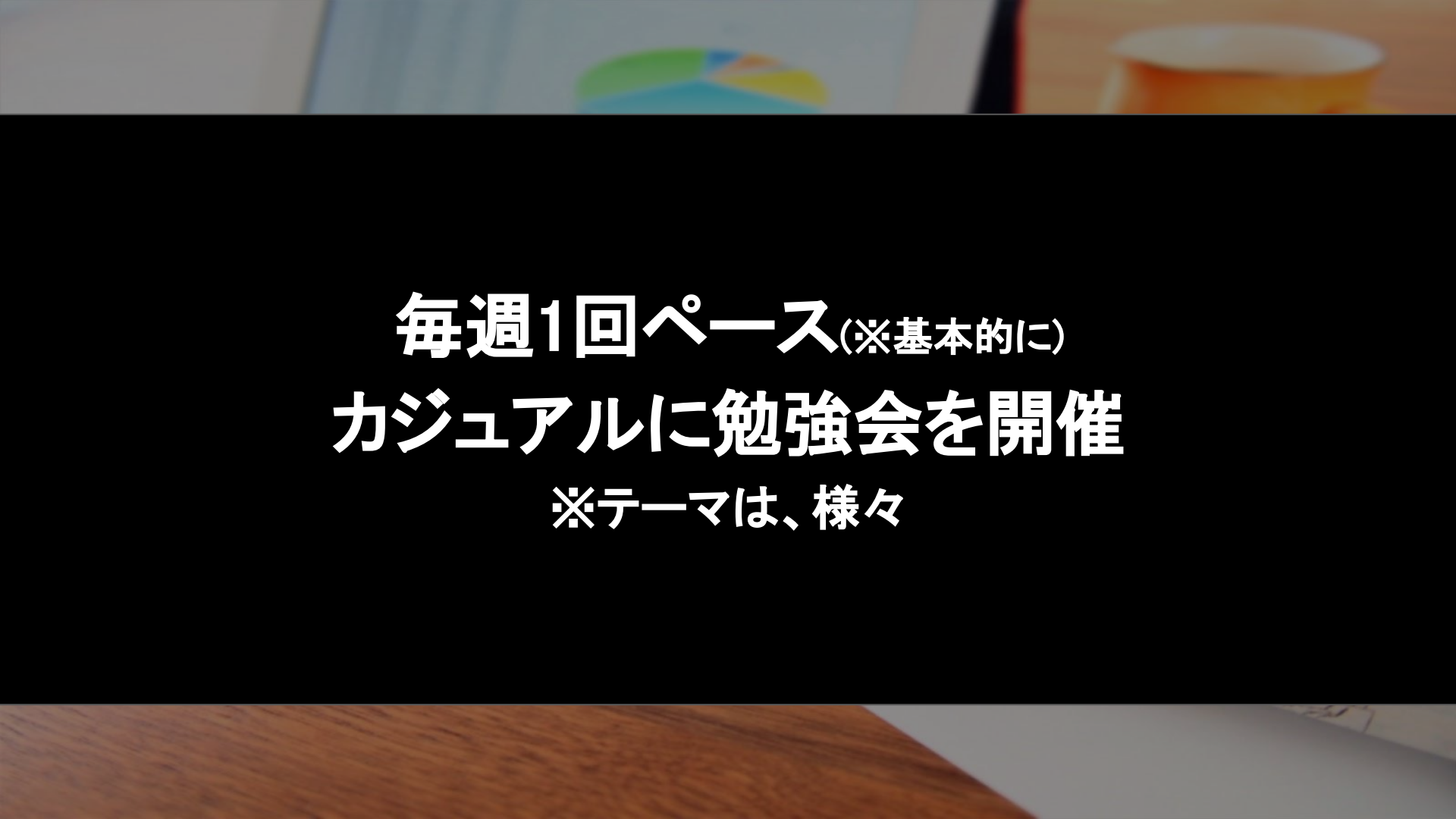
宇宙で最も力強いのは
幅広い興味である

理論物理学者
アルベルト・アインシュタイン



勉強したいと思う源泉は、新しいもの、珍しいもの、自分とは違うものに対する好奇心です

ファーストリテイリング 代表取締役会長兼社長 柳井正



毎週1回ペース(※基本的に)
カジュアルに勉強会を開催
※テーマは、様々

A desk setup featuring a laptop with a red case displaying a pie chart, a yellow mug, a red and white patterned pen, and an open book with French text. The scene is set on a wooden desk.

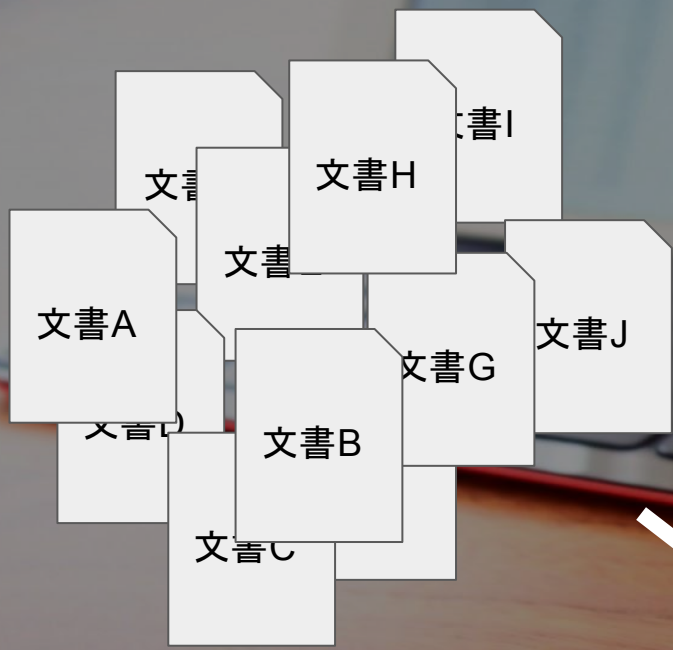
今回は... 自然言語処理とトピックモデル

A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a red pen with a gold clip, and an open book with French text. The background is a wooden desk.

トピックモデルを学ぶ上で、
自然言語処理は欠かせない？

A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a red and white patterned pen, and an open book with French text. The background is a wooden desk.

自然言語処理の1分野に トピックモデルがある(もしくは、あった)



トピックを抽出

国会
首相
内閣
衆議院
選挙
...

政治

選手
ゴール
ボール
試合
球場
...

スポーツ

病院
薬
健康
手術
難病
...

医療

...



それぞれの文書がもつトピックを推定

サッカー協会は代
表選考のため強化
合宿を...

スポーツ

次期衆議院選挙に
向け与党は...

政治

難病指定について
厚生労働省は...

医療+政治



トピックを抽出

トピックモデルによるデータ解析

大量のデータを活用し、有益な情報を発見するためのツールとしてトピックモデル(topic model)が注目されている

スポーツ

政治

医療＋政治



トピックを抽出

**トピックモデルを用いることで
人手を介在させることなく、大量の文書集合から話題と
なっているトピックを抽出することが可能**

スポーツ

政治

医療＋政治



トピックを抽出

トピックモデルを用いることで
それぞれの文書がどのようなトピックを持っているか
トピックの近い文書を探索する、文書を分類することも可能

スポーツ

政治

医療＋政治

トピックモデル応用分野

画像処理

推薦システム

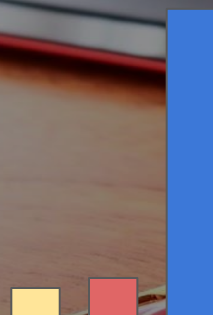
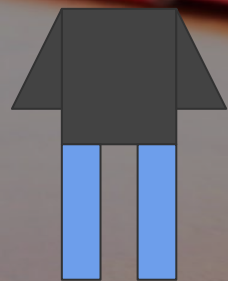
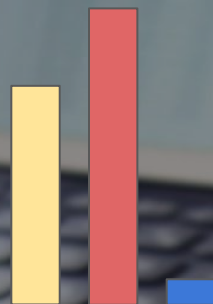
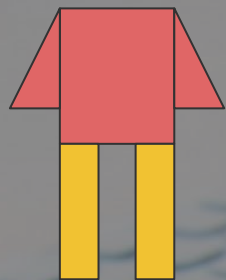
ソーシャルネットワーク

バイオインフォマティクス

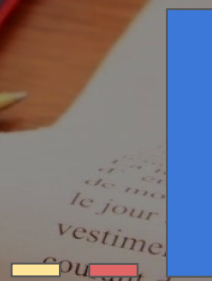
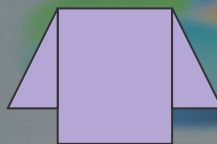
音楽情報処理

予測・異常検知

サンプル画像



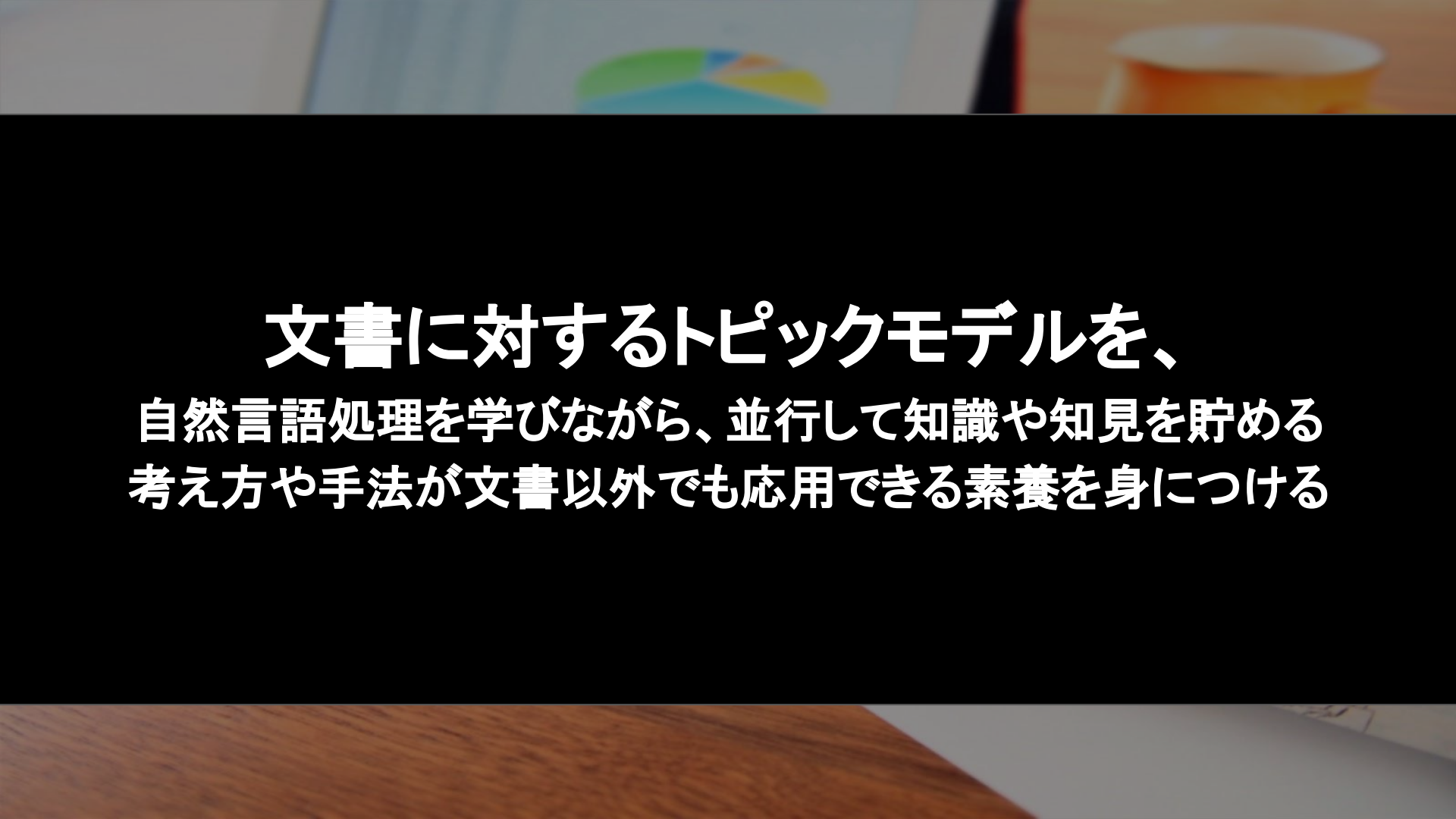
上衣写真



下衣写真



例えば、コーディネート推薦
文書ベクトルも、画像ベクトルも同じだよね



**文書に対するトピックモデルを、
自然言語処理を学びながら、並行して知識や知見を貯める
考え方や手法が文書以外でも応用できる素養を身につける**

ベースとなる参考文献・図書

- [1]情報抽出・固有表現抽出のための基礎知識(近代科学社)
- [2]言語処理システムをつくる(近代科学社)
- [3]深層学習による自然言語処理(講談社)
- [4]トピックモデル(講談社)
- [5]スパース性に基づく機械学習(講談社)
- [6]オライリー自然言語処理
- [7]画像/言語同時埋め込みベクトル空間の構築に向けた埋め込み
粒度比較検討(東北大学・理化学研究所)
- [8]前方文脈の埋め込みを利用して日本語述語項構造解析
(東北大学 乾・鈴木研究室)



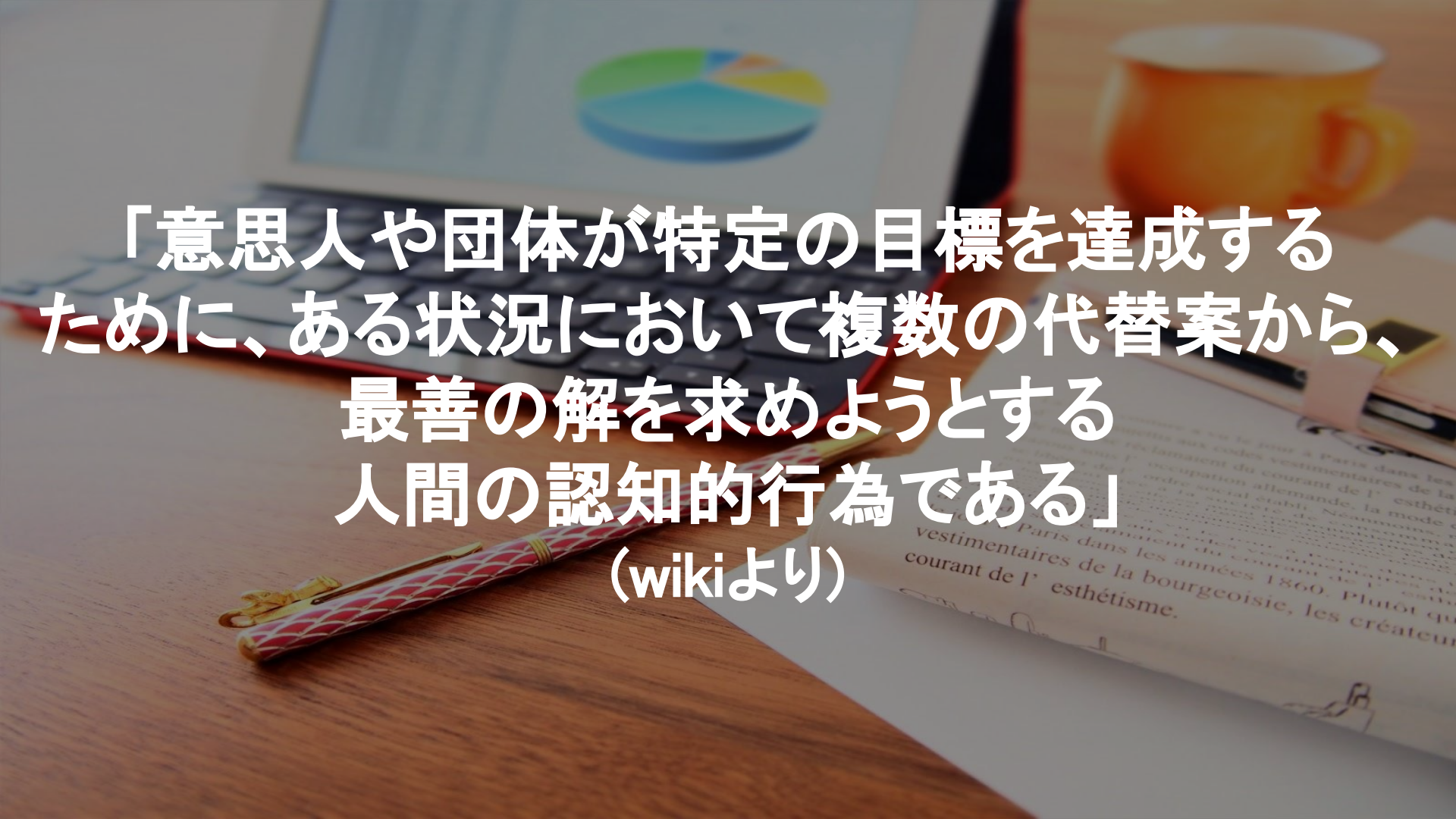
Google Colaboratory

Pythonで自然言語処理 & トピックモデルを学ぶ

第1回目

意思決定とは

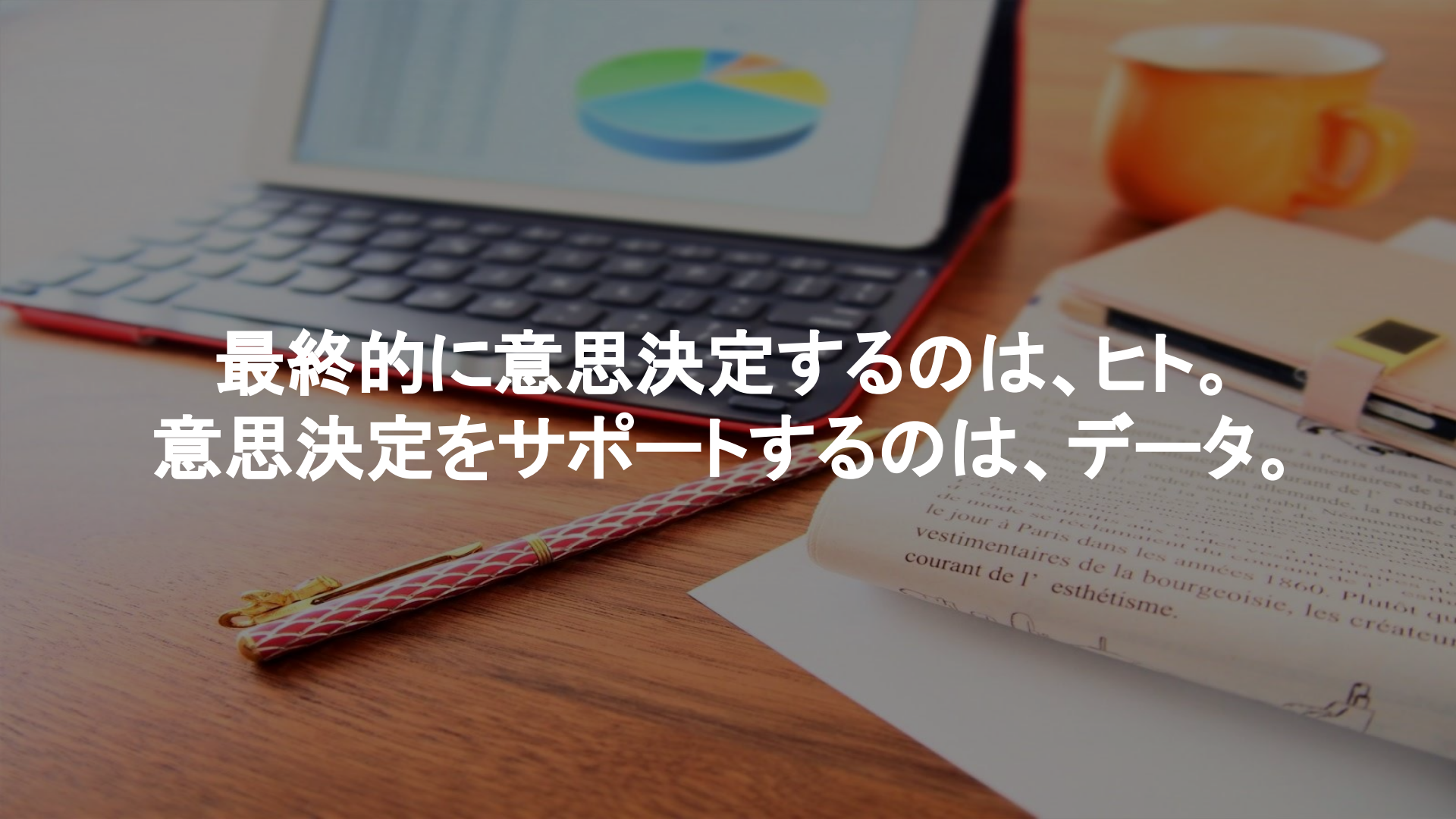


A desk setup featuring a laptop with a pie chart on its screen, a yellow cup of orange juice, a red and white patterned pen, and a document with French text. The text is overlaid in large white characters.

「意思人や団体が特定の目標を達成する
ために、ある状況において複数の代替案から、
最善の解を求めようとする
人間の認知的行為である」
(wikiより)

A desk setup featuring a laptop with a pie chart on its screen, a red and gold patterned pen, a notebook with a yellow clip, and a yellow cup. The background is a wooden desk.

データで意思決定するとは？

A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a red and white patterned pen, and an open book with French text. The text is overlaid in large white characters.

最終的に意思決定するのは、ヒト。
意思決定をサポートするのは、データ。

A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a red and white patterned pen, and an open book with French text. The scene is softly lit, creating a professional yet cozy atmosphere.

データから意思決定するまでの プロセスやアプローチ

1. 收集



La haute couture a vu le jour à Paris dans les
d' une aristocratie aux codes vestimentaires de la
de mode se réclamaient du courant de l' esthétisme
surtout sous l' occupation allemande, la mode c
se libéra de l' ordre social établi. Néanmoins, r
accréditant l' idée d' la société de consommation
l' individualisme, cette société de consommation
de la haute couture, pour l' aristocratie, la bourgeoisie
d' être assujettis aux codes vestimentaires de l' est
de mode se réclamaient du courant de l' esthétisme
le jour à Paris dans les années 1860. Plutôt qu
vestimentaires de la bourgeoisie, les créateurs
courant de l' esthétisme.

1.
收集



2.
整形



1.
収集

2.
整形



3.
集計

Le jour à Paris dans les années 1860. Plutôt que vestimentaires de la bourgeoisie, les créateurs de mode se réclamaient du courant de l'esthétisme.

1.
収集

2.
整形

3.
集計



4.
可視化

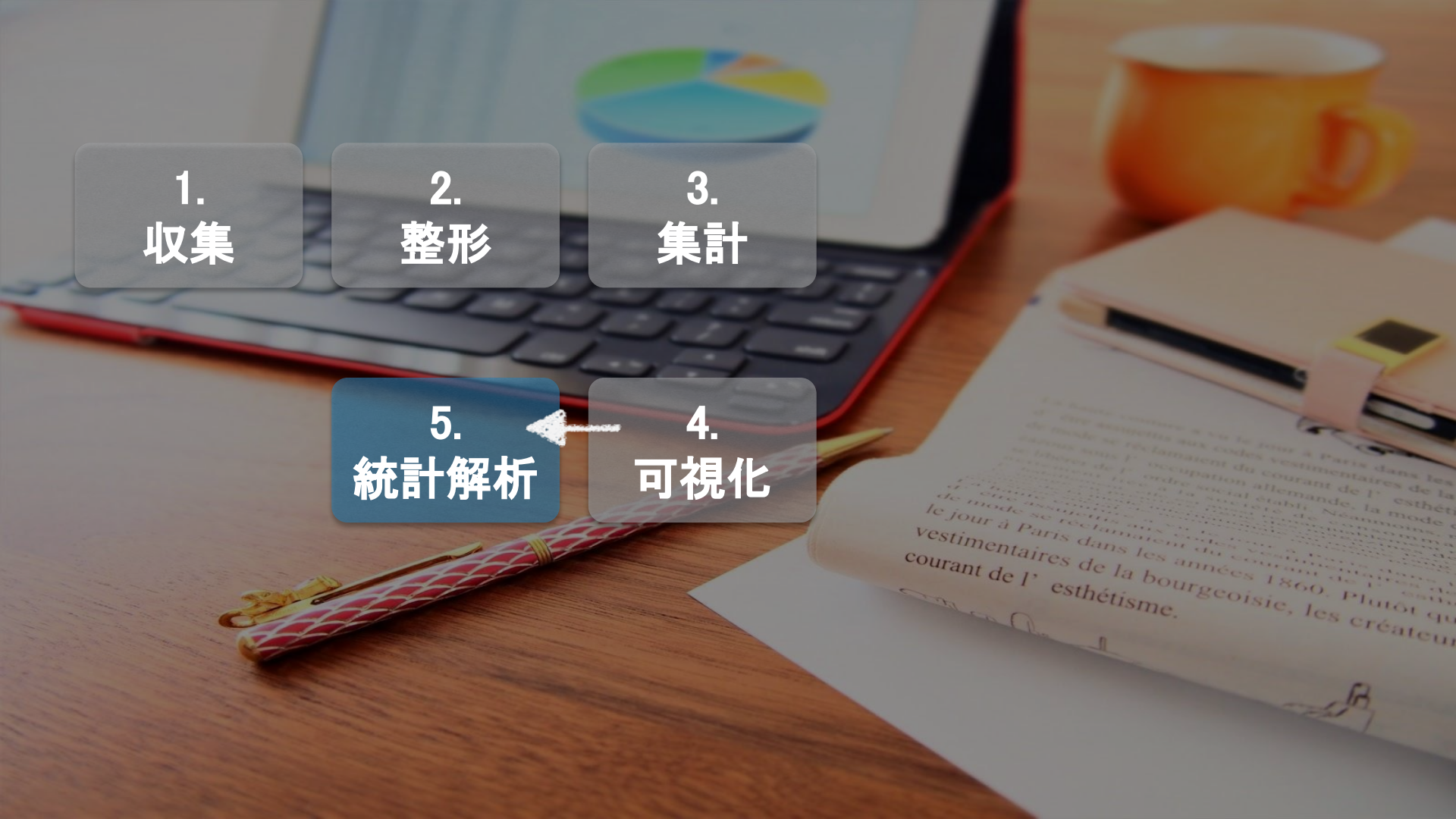
1.
収集

2.
整形

3.
集計

5.
統計解析

4.
可視化



1.
収集

2.
整形

3.
集計

5.
統計解析

4.
可視化

6.
機械学習



1.
収集

2.
整形

3.
集計

5.
統計解析

4.
可視化

6.
機械学習

↓
7.
考察

1.
収集

2.
整形

3.
集計

5.
統計解析

4.
可視化

6.
機械学習

7.
考察

8.
意思決定



プロセス

1.
収集

2.
整形

3.
集計

5.
統計解析

4.
可視化

6.
機械学習

7.
考察

8.
意思決定

データエンジニア

1.
収集

2.
整形

3.
集計

MySQL®

SQLite

hadoop



5.
統計解析

4.
可視化

6.
機械学習



DEEP
LEARNING
INSTITUTE

7.
考察

8.
意思決定

1.
収集

2.
整形

3.
集計

5.
統計解析

4.
可視化

6.
機械学習

7.
考察

8.
意思決定



データアナリスト



1.
収集

2.
整形

3.
集計

5.
統計解析

4.
可視化

6.
機械学習

7.
考察

8.
意思決定

K Keras



AIエンジニア



pandas

PyTorch



1.
収集

2.
整形

3.
集計

5.
統計解析

4.
可視化

6.
機械学習

意思決定者
(現場のヒト)

7.
考察

8.
意思決定

1.
収集

2.
整形

3.
集計

5.
統計解析

4.
可視化

6.
機械学習

データサイエンティスト
(意思決定のサポーター)

7.
考察

8.
意思決定

A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a red and white patterned pen, and an open book with French text. The background is a wooden desk.

意思決定の代表的な分野 自然言語処理 & トピックモデル

自然言語処理への入り口



A desk setup featuring a laptop with a pie chart on its screen, a yellow cup, a red and white patterned pen, and an open book with French text. The scene is set on a wooden desk.

必要な基本知識を
初回紹介していきます

扱う言語・文法
(日本語, 英語, ...)

言語処理の技術
(形態素分析, 構文解析法...)

ライブラリ・辞書
(NLTK, Mecab...)

機械学習 前処理
(ゼロパディング, ストップワード,
コーパスクリーニング,...)

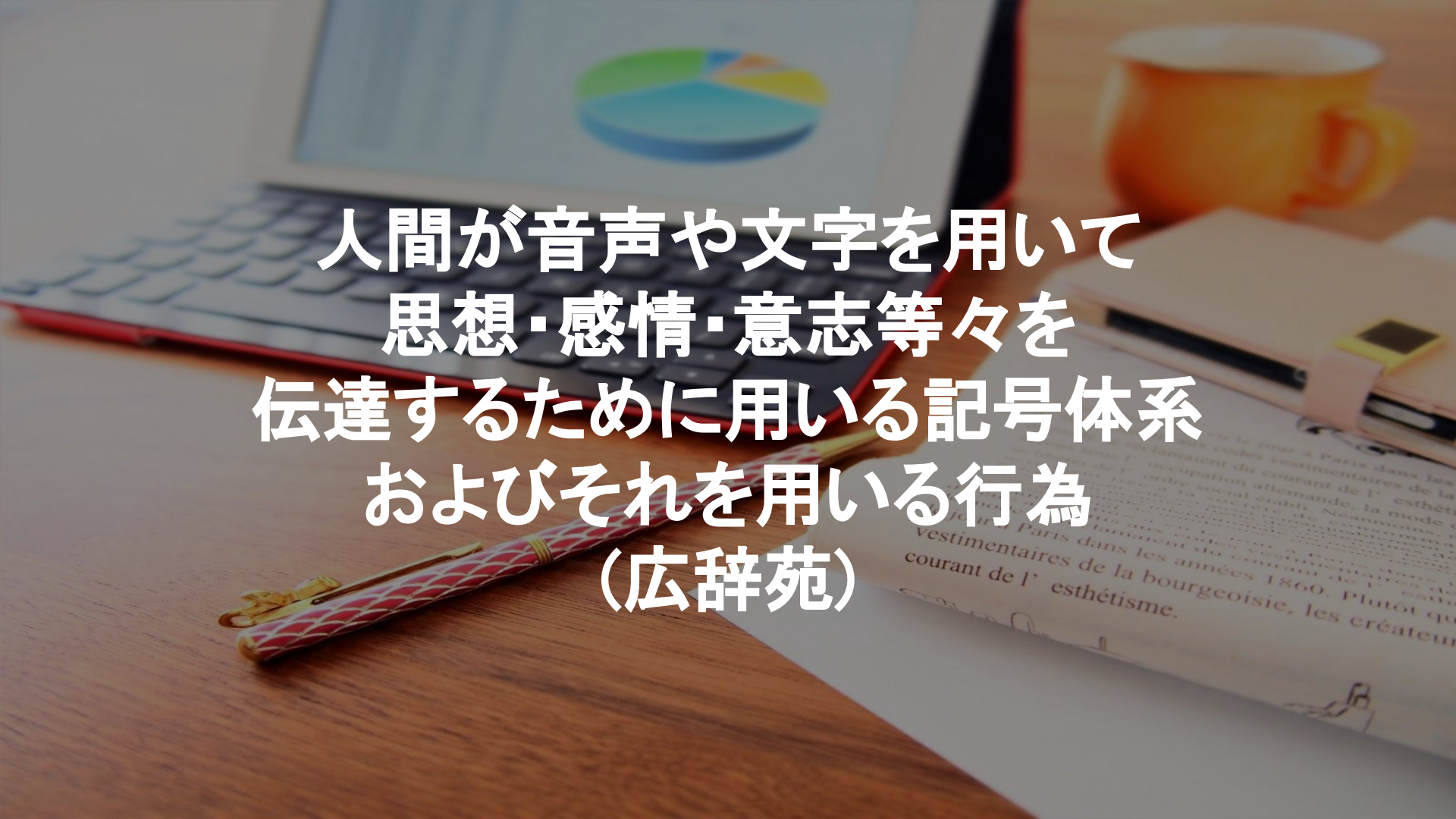
扱う言語・文法 (日本語, 英語, ...)



A desk setup featuring a laptop with a red case displaying a pie chart, a red and white patterned pen, a notebook with a yellow cover, and a yellow cup. The background is a wooden desk.

言語 (Language)

Le jour à Paris dans les années 1860. Plutôt que vestimentaires de la bourgeoisie, les créateurs de mode se réclamaient du courant de l'esthétisme.

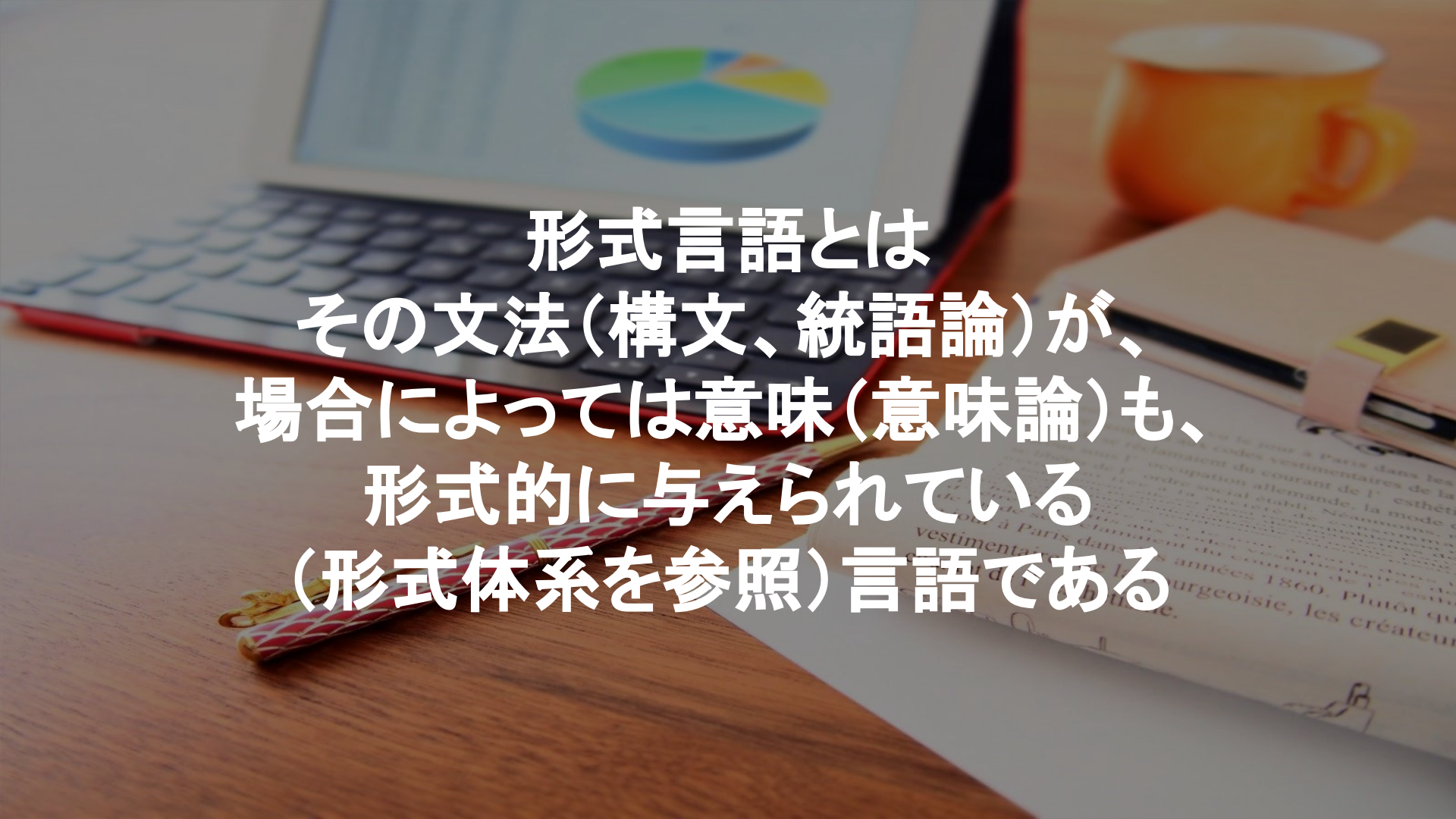
A blurred background image of a desk. In the upper left, a laptop screen shows a pie chart. To the right is a yellow mug. In the foreground, a red and white patterned pen lies on a wooden surface. Below the pen is an open book with French text. The text is overlaid in large white characters.

人間が音声や文字を用いて
思想・感情・意志等々を
伝達するために用いる記号体系
およびそれを用いる行為
(広辞苑)

音声や文字によって、
人の意志・思想・感情などの情報を
表現したり伝達する、あるいは
他者のそれを受け入れ、
理解するための約束・規則および、
そうした記号の体系
(大辞泉)

形式言語と自然言語



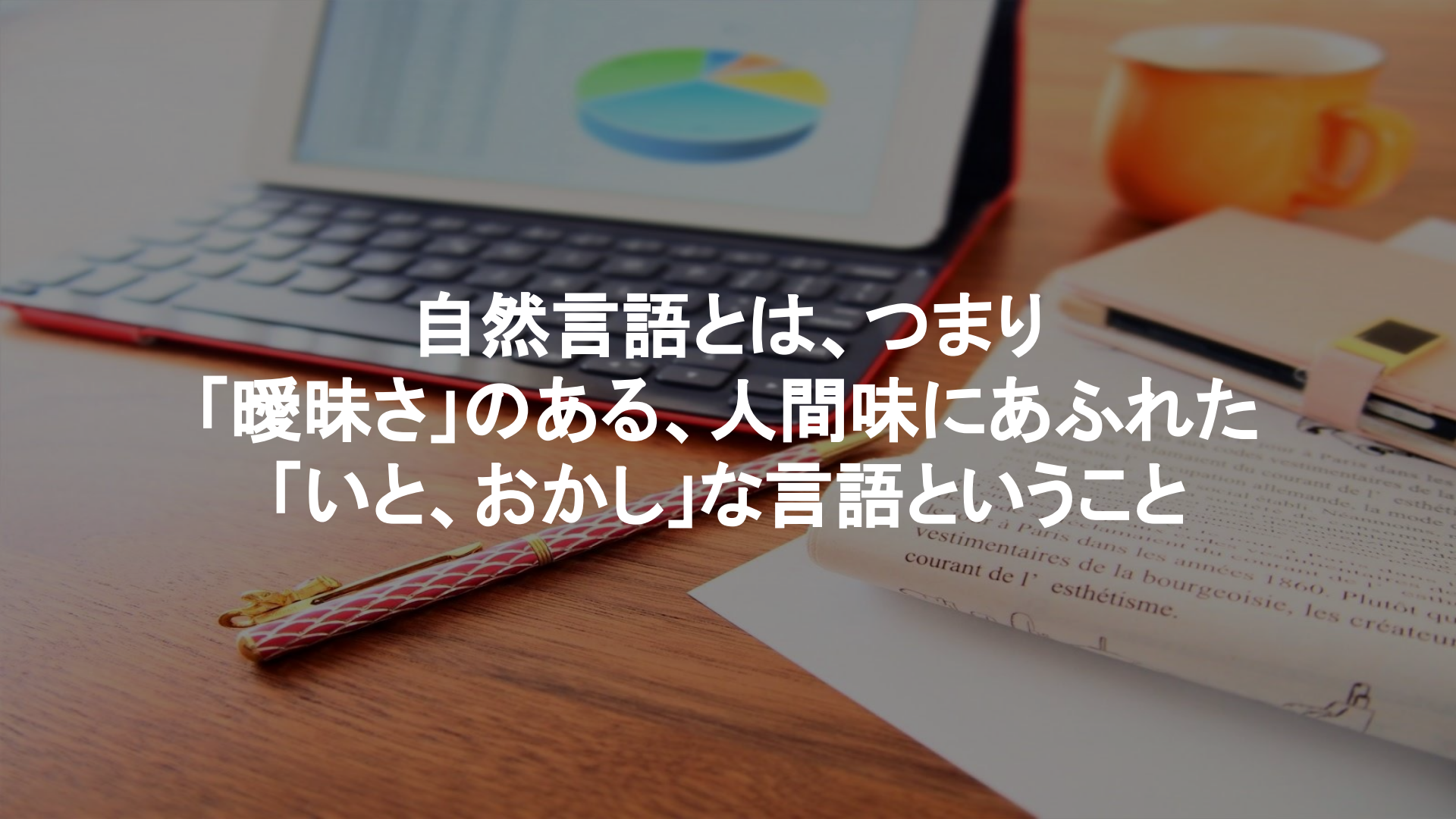
A blurred background image of a desk. On the left, a laptop is open, showing a pie chart on its screen. To the right of the laptop is a yellow mug. In the foreground, there is a notebook with a yellow cover and a red pen resting on it. The text is overlaid on this background.

形式言語とは
その文法(構文、統語論)が、
場合によっては意味(意味論)も、
形式的に与えられている
(形式体系を参照)言語である

A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a red pen, and a book with a yellow cover. The text is overlaid on the laptop and the book.

形式言語とは、つまり
「曖昧さ」のない、意味が明確な言語。
プログラミングでは、オートマトン
「機械言語」としての立ち位置

自然言語とは、形式言語と対比され
人間によって日常の意思疎通のために
用いられる、文化的背景を持って
自然に発展してきた言語。
人間がお互いにコミュニケーションを行うため
の自然発生的な言語である。

A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a red pen with a gold clip, and a book with French text. The text is overlaid in the center.

自然言語とは、つまり
「曖昧さ」のある、人間味にあふれた
「いと、おかし」な言語ということ

言語

```
graph TD; A[言語] --> B[形式言語]; A --> C[自然言語]; B --> D[Java]; B --> E[Fortran]; B --> F[C]; B --> G[Python]; B --> H[PHP]; C --> I[英語]; C --> J[中国語]; C --> K[日本語]; C --> L[ロシア語];
```

形式言語

Java

Fortran

C

Python

PHP

自然言語

日本語

英語

中国語

ロシア語

A desk setup featuring a laptop with a pie chart on its screen, a red and gold patterned pen, and an open book with French text. A green oval overlay contains the title text.

言語処理の技術 (形態素分析, 構文解析法…)

構文と形態素

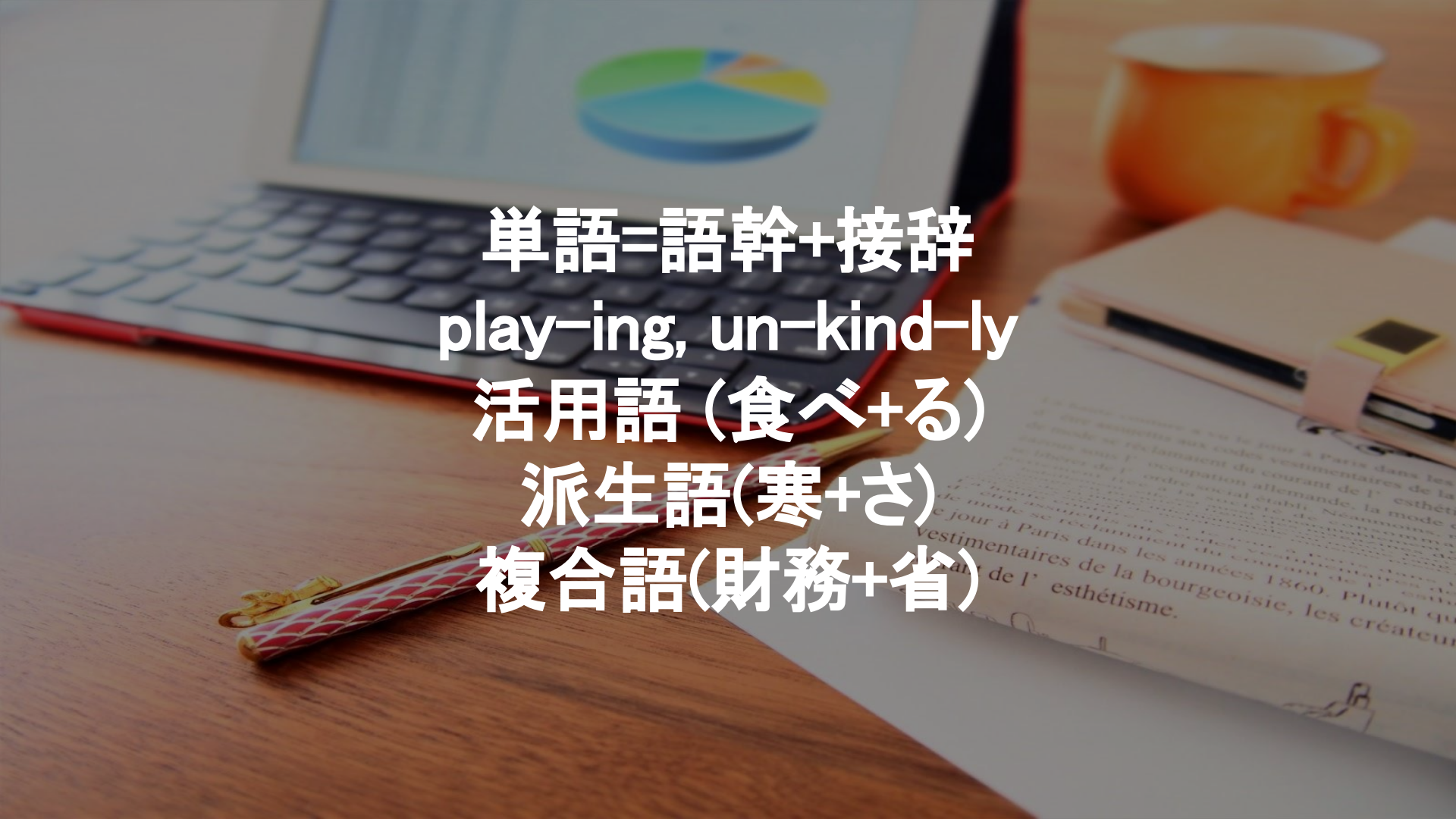


A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a red and white patterned pen, and an open book with French text. The background is a wooden desk.

構文 文(Sentence)の、 構成(Structure)

A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a red pen with a gold clip, and an open book with French text. The text is overlaid in the center of the image.

形態素
意味を持つ最小の言語単位
単語よりも小さい単位

A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a notebook with a yellow tab, and a red and white patterned pen. The background is a wooden desk.

単語=語幹+接辞
play-ing, un-kind-ly
活用語(食べ+る)
派生語(寒+さ)
複合語(財務+省)

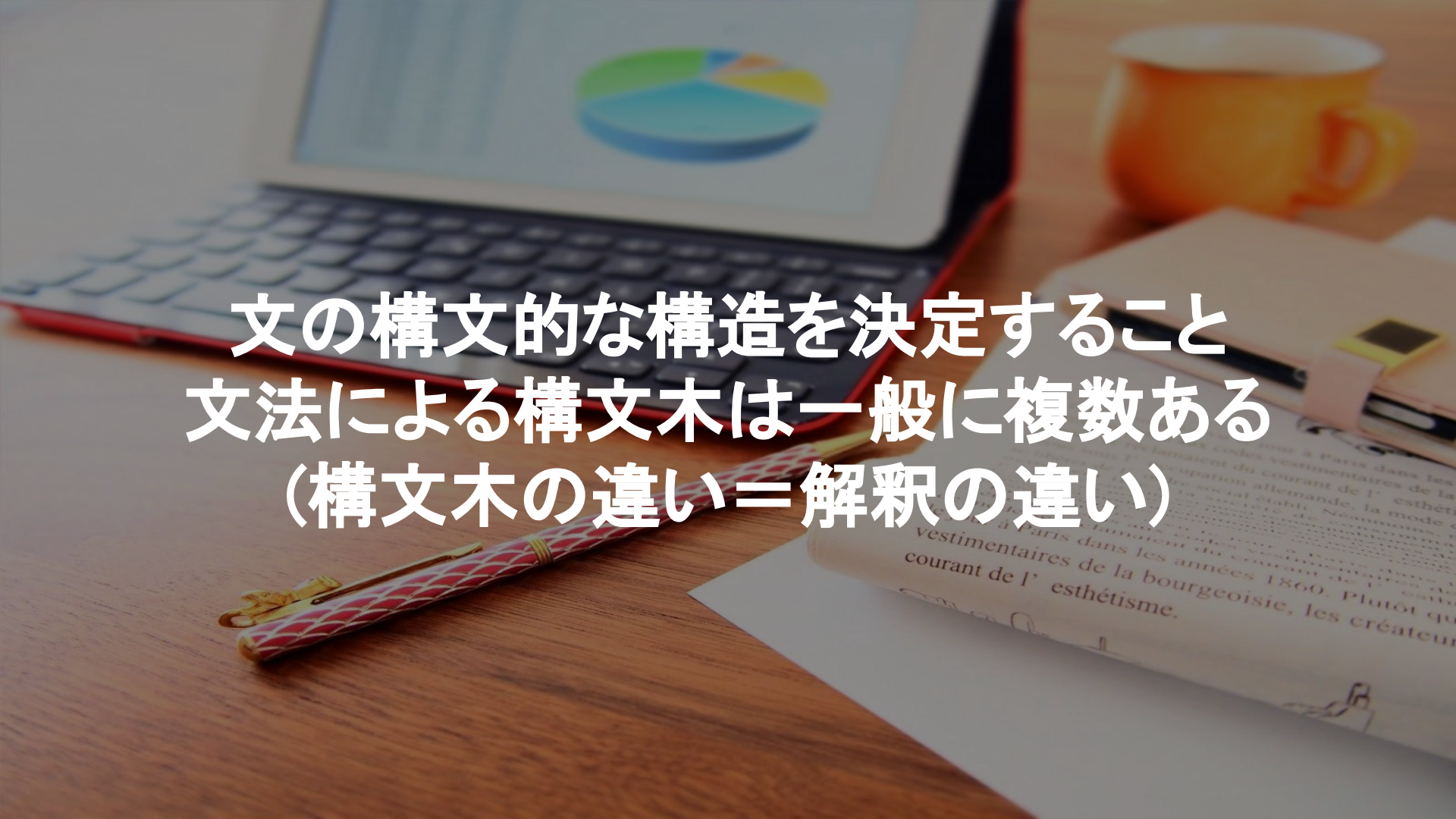
形態素分析



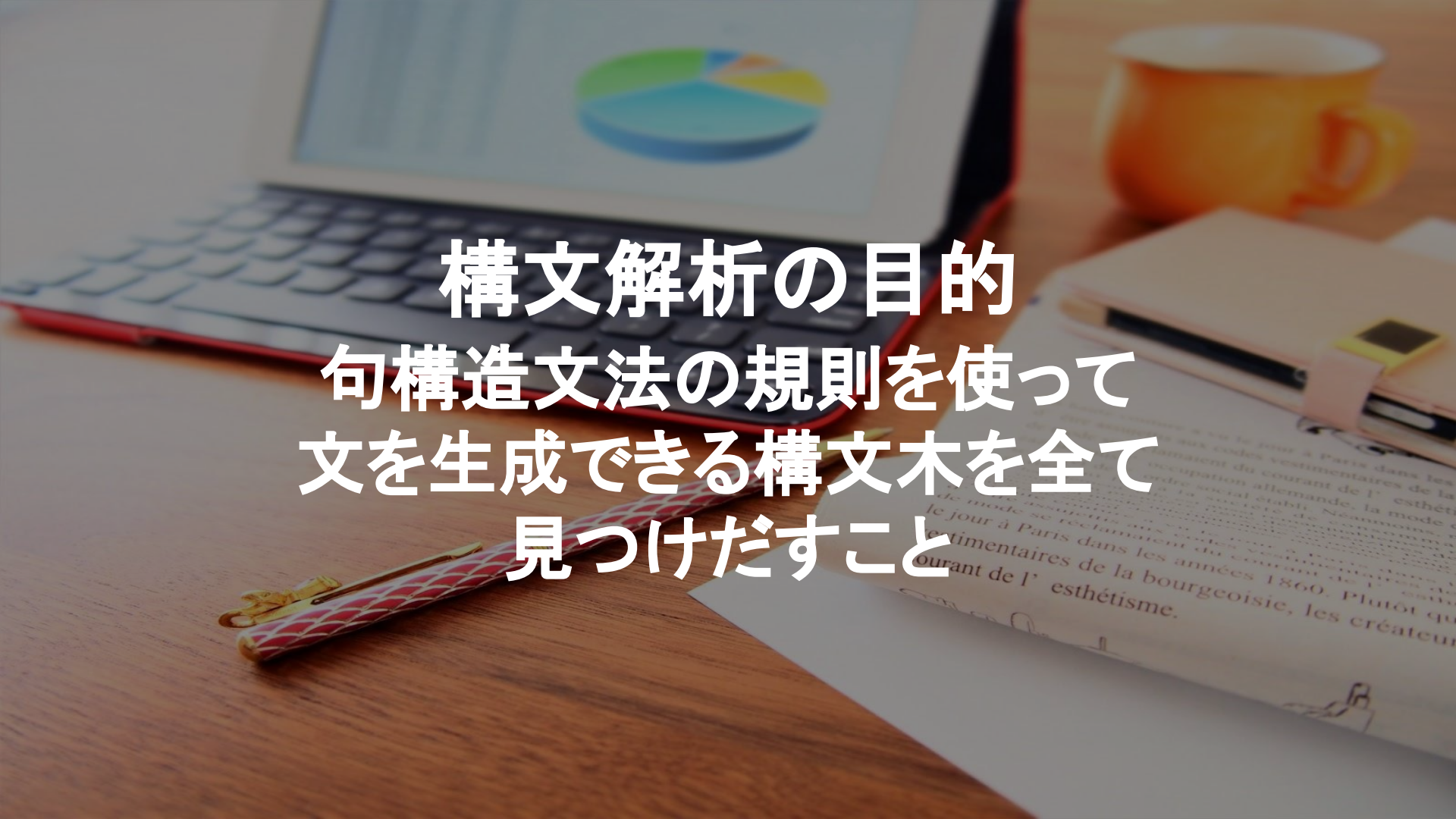
構文解析の前に行われる処理

- ・形態素区切りを決める
- ・品詞を決める
- ・単語境界を決める

A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a red and white patterned pen, and an open book with French text. The text on the book includes "le jour à Paris dans les années 1860. Plutôt qu'...", "vestimentaires de la bourgeoisie, les créateurs...", and "courant de l'esthétisme." The title "構文解析" is overlaid in the center.

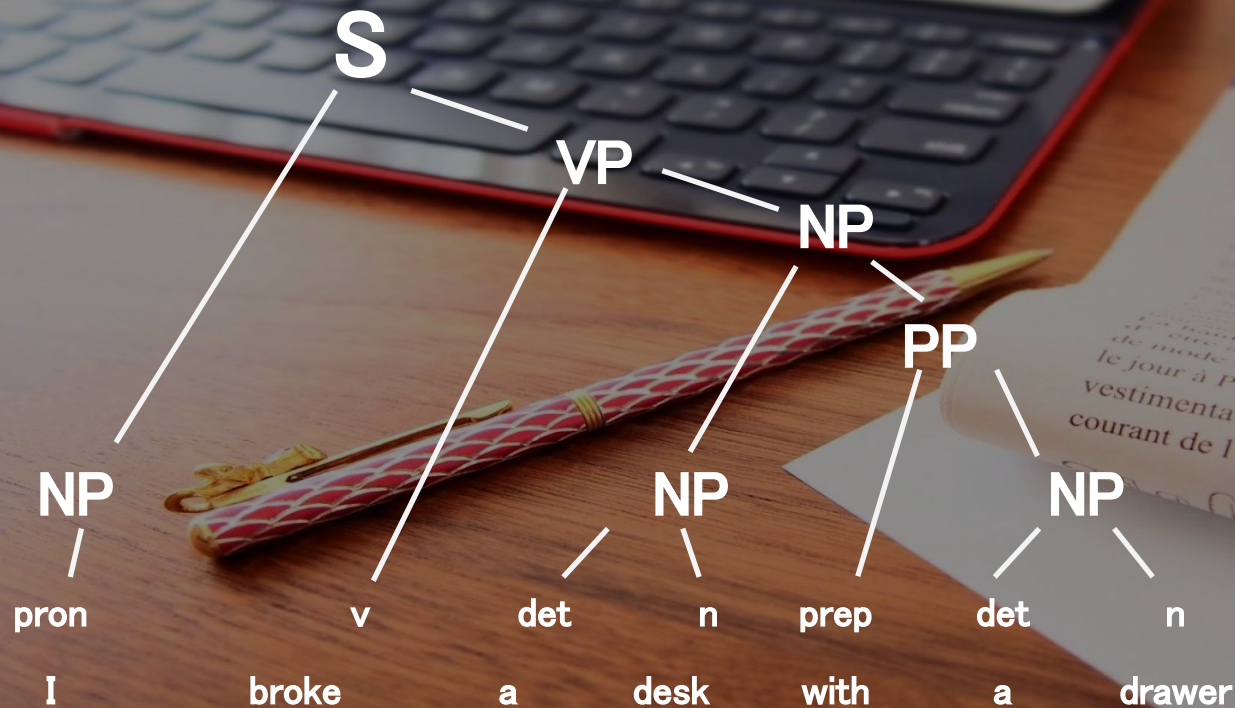
A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a red pen, and a book. The text is overlaid on the image.

文の構文的な構造を決定すること
文法による構文木は一般に複数ある
(構文木の違い＝解釈の違い)

A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a notebook, and a red pen. The text is overlaid on the laptop and notebook.

構文解析の目的
句構造文法の規則を使って
文を生成できる構文木を全て
見つけたること

構文解析、1例



句構造規則

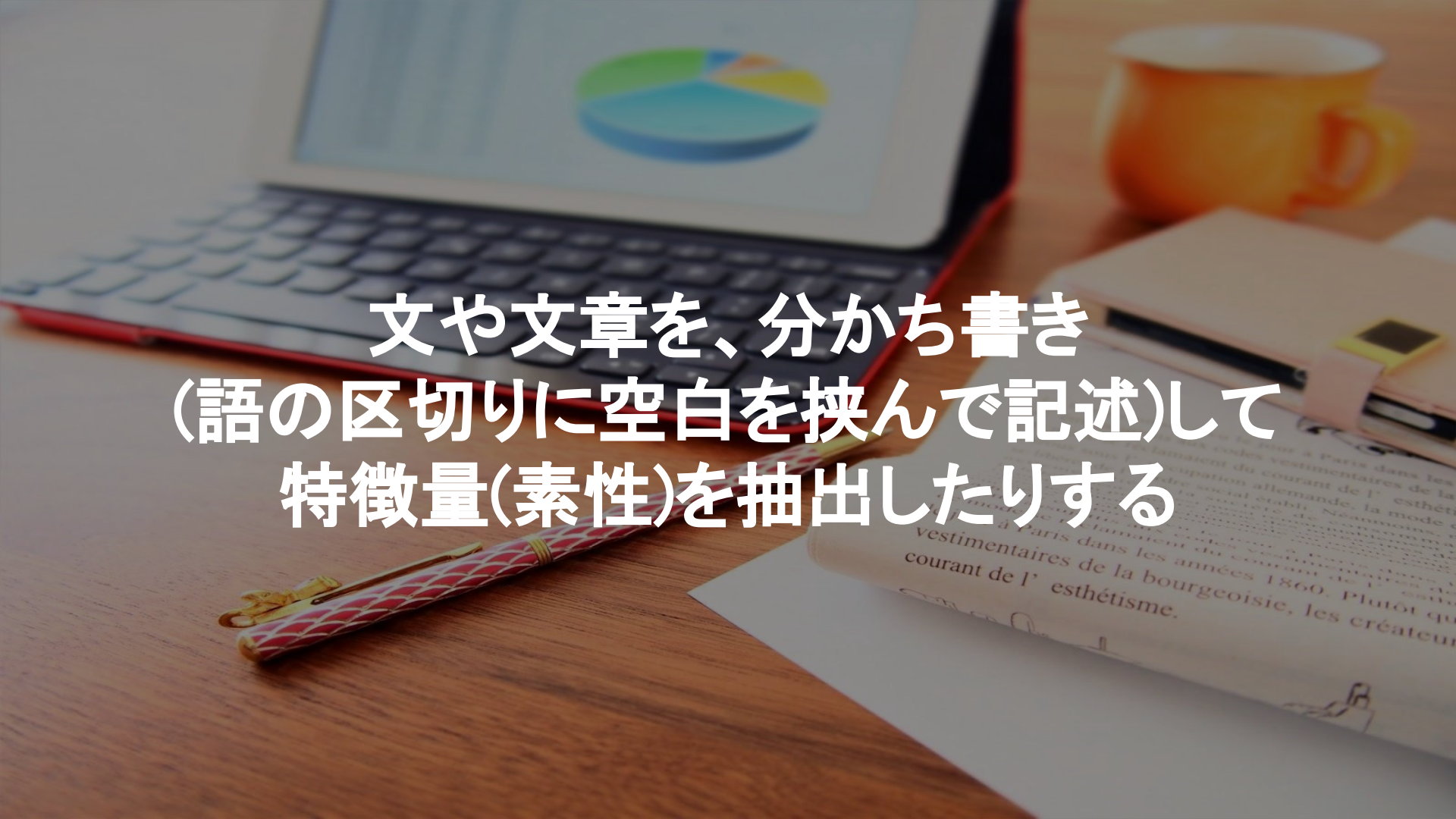
辞書規則

素性と単一化



素性構造(feature structure)の構成要素で
属性と属性値のペアからなるものを
素性という

文法(原理・規則)に基づき、
2つの素性構造を素性と値に矛盾がない
ように1つの素性構造にまとめる操作を
単一化という

A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a red pen, and a book. The text is overlaid on the laptop and pen.

文や文章を、分かち書き
(語の区切りに空白を挟んで記述)して
特徴量(素性)を抽出したりする

A desk setup featuring a laptop with a red case displaying a pie chart, a yellow cup, a notebook with a yellow cover, and a red and white patterned pen. A yellow oval overlay contains the text.

ライブラリ・辞書 (NLTK, Mecab...)

A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a red and white patterned pen, and a book with French text. The text is overlaid in the center.

NLTK

英語周辺のコーパスを用意してくれる オープンソースツール

A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a red pen with a gold clip, and an open book with French text. The background is a wooden desk.

MeCab

日本語のコーパスを用意してくれる オープンソースツール

A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a notebook with a yellow cover, and a red and white patterned pen. The background is a wooden desk.

NLTK, MeCab

これらがあることで基本的な構造分析、
形態素分析、素性や単一化を行う必要が
なくなる(手間が省ける)

A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a notebook with a yellow cover, and a red and white patterned pen. The background is a wooden desk.

Janome

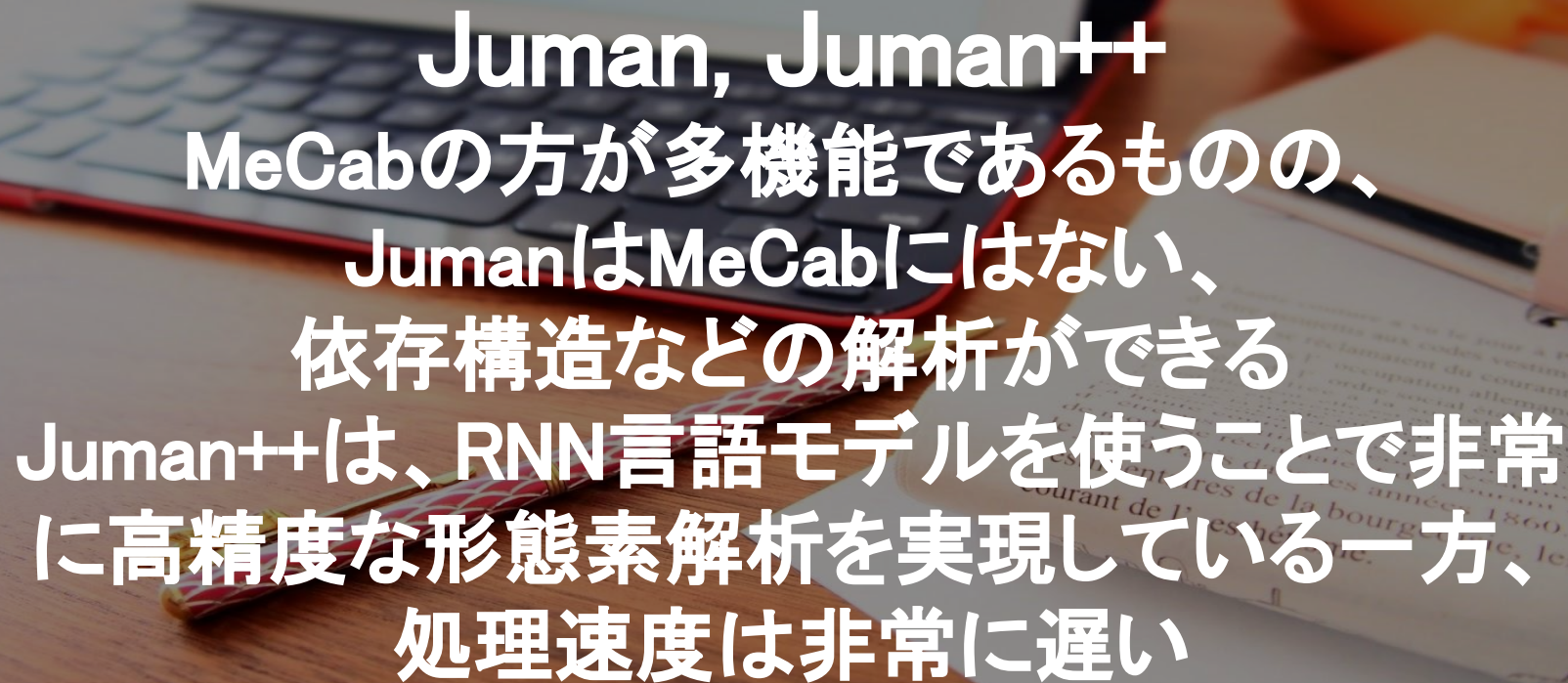
MeCabと同じように、日本語の形態素
を得意としているものであるが...
MeCabの方が処理が早い！

NEologd

MeCabと同じように、日本語の形態素
を得意としているものであるが...

MeCabの方が処理が早い！

一方で、Web上のあらゆる新語が追加された巨
大な辞書なので、
Twitter分析でよく使われる



**Juman, Juman++
MeCabの方が多機能であるものの、
JumanはMeCabにはない、
依存構造などの解析ができる
Juman++は、RNN言語モデルを使うことで非常に
高精度な形態素解析を実現している一方、
処理速度は非常に遅い**

A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a red pen, and a book. The text is overlaid on the laptop and the book.

Unidic

国立教育政策研究所が作成した辞書
実用的には、形態素が短すぎるため、
扱いづらい

Sudachi

商用利用を目的として、徳島からランチ
された...企業の支援が手厚く保守性が高い

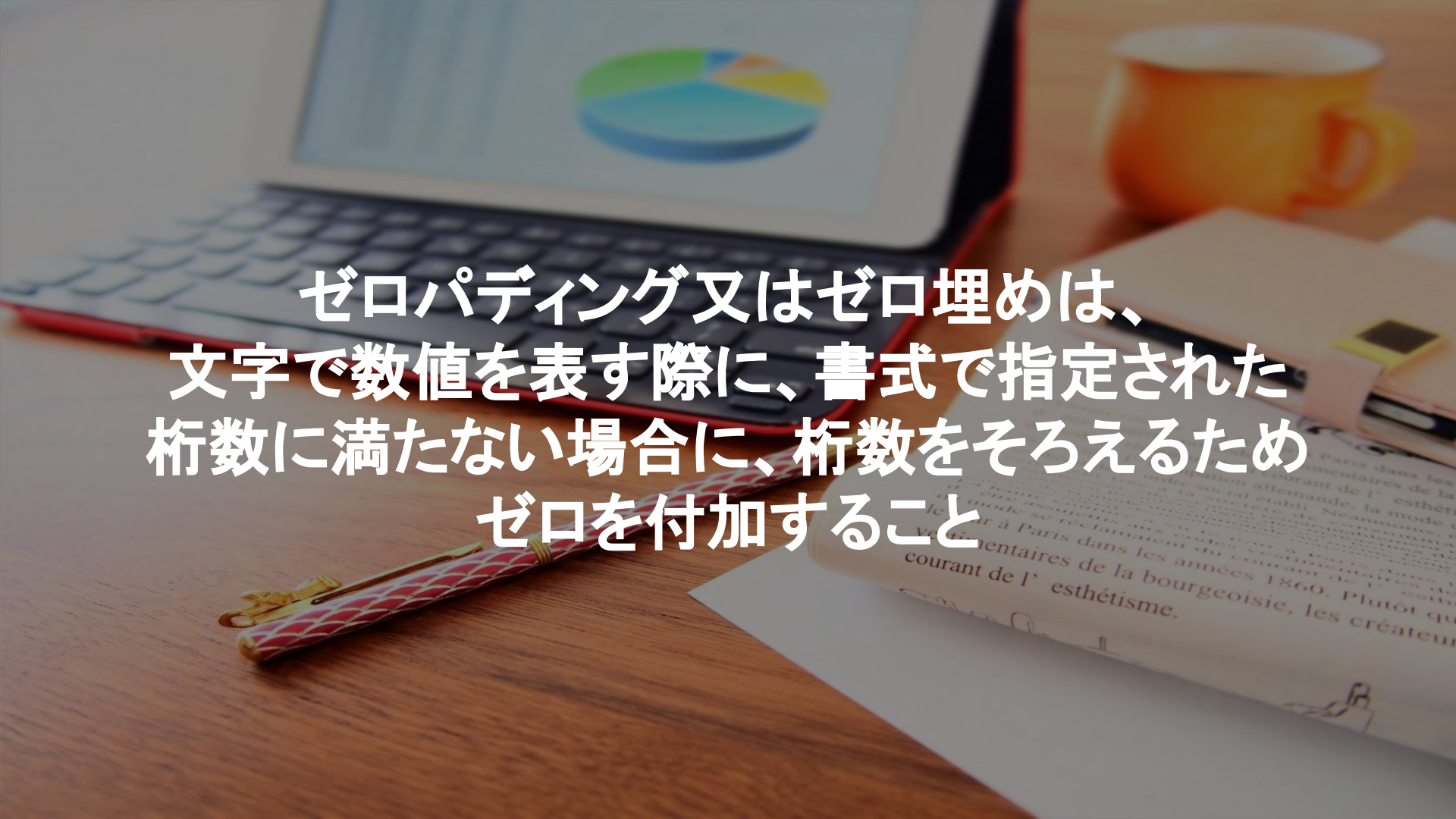
A単位(Unidic相当)B単位(IPAdic相当)

C単位(NEologd相当)という分割長さにも対応



機械学習 前処理 (ゼロパディング, ストップワード, コーパスクリーニング,...)

ゼロパディング

A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a red pen, and a book. The text is overlaid on the center of the image.

ゼロパディング又はゼロ埋めは、
文字で数値を表す際に、書式で指定された
桁数に満たない場合に、桁数をそろえるため
ゼロを付加すること

A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a red and white patterned pen, and an open book with French text. The scene is set on a wooden desk.

機械学習モデルのデータ前処理や 自然言語処理でよく扱う

ストップワード

Le jour à Paris dans les années 1860. Plutôt que vestimentaires de la bourgeoisie, les créateurs de mode se réclamaient du courant de l'esthétisme.

A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a red pen with a gold clip, and a book with French text. The text is overlaid in white.

MeCabとかNLTKを利用しても
やはり不要な言葉を抽出してしまう場面で
不要な言葉をあらかじめ用意しておく

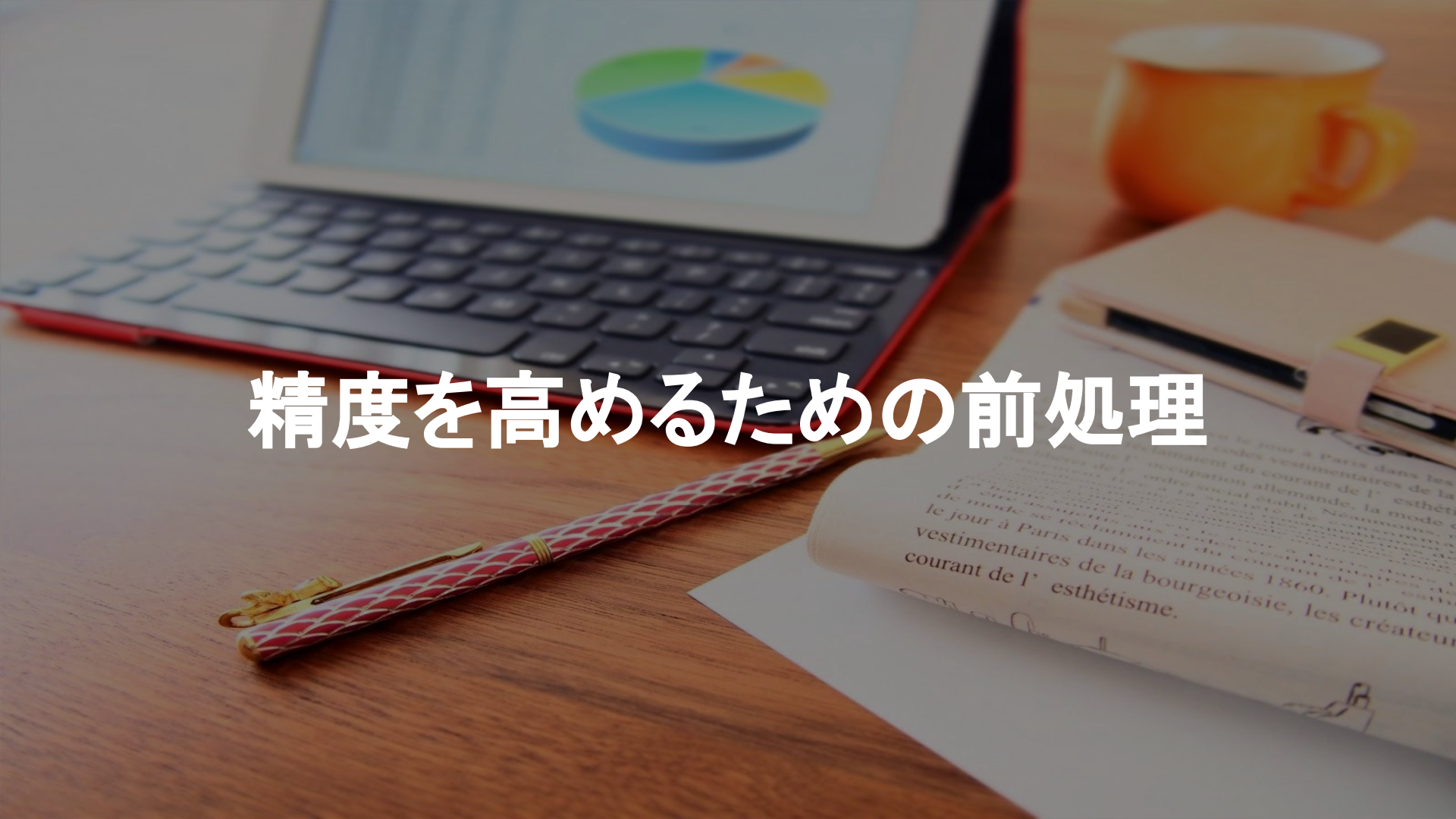
A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a red and white patterned pen, and an open book with French text. The background is a wooden desk.

ストップワードを適用することで
より正確な情報、単語を抜き出せる

コーパスクリーニング



精度を高めるための前処理



コーパスクリーニング 1例

- ・大文字/小文字の変換
- ・句読点の統一
- ・用語(同義語)の統一
- ・テキスト属性によるクリーニング
- ・無駄な行やスペースの削除
- ・SNSテキスト特有の文字(列)
- ・半角/全角の統一(正規化)

..., etc

扱う言語・文法
(日本語, 英語, ...)

言語処理の技術
(形態素分析, 構文解析法...)

ライブラリ・辞書
(NLTK, Mecab...)

機械学習 前処理
(ゼロパディング, ストップワード,
コーパスクリーニング,...)

扱う言語・文法
(日本語, 英語, ...)

言語処理の技術

(形態素分析, 構文解析法...)

これら一連の作業をコンピュータで行う
→ **自然言語処理**

(NLTK, Mecab...)

機械学習 前処理

(ゼロパディング, ストップワード,
コーパスクリーニング,...)

A desk setup featuring a laptop with a red case displaying a pie chart, a red and white patterned pen, a notebook with a yellow cover, and a yellow cup of coffee. The background is a wooden desk.

自然言語処理 (Natural Language Processing)

A desk setup featuring a laptop with a pie chart on its screen, a yellow mug, a red and white patterned pen, and an open book with French text. The background is a warm, wooden desk.

計算機で自然言語を 「処理」すること 人工知能の研究の一分野

A desk setup featuring a laptop with a red case displaying a 3D pie chart, a yellow mug, a red and white patterned pen, and a book with French text. The text is overlaid in the center.

トピックモデル
こちらは次回以降、紹介していきます