

【公報種別】特許法第17条の2の規定による補正の掲載

【部門区分】第6部門第3区分

【発行日】平成20年5月22日(2008.5.22)

【公開番号】特開2005-302023(P2005-302023A)

【公開日】平成17年10月27日(2005.10.27)

【年通号数】公開・登録公報2005-042

【出願番号】特願2005-110069(P2005-110069)

【国際特許分類】

G 06 F 17/21 (2006.01)

【F I】

G 06 F 17/21 5 9 2 D

【手続補正書】

【提出日】平成20年4月4日(2008.4.4)

【手続補正1】

【補正対象書類名】特許請求の範囲

【補正対象項目名】全文

【補正方法】変更

【補正の内容】

【特許請求の範囲】

【請求項1】

テキストを自動的に大文字化するためのキャピタライゼーションモデルをトレーニングする方法であって、

トレーニング文書が特定のユーザに関連付けられていることを必要とする制限事項を満たす前記トレーニング文書を収集するステップと、

前記収集したトレーニング文書を用いて前記キャピタライゼーションモデルをトレーニングするステップと、

を含み、

前記収集したトレーニング文書を用いて前記キャピタライゼーションモデルを該トレーニングするステップは、前記収集したトレーニング文書中のキャピタライゼーションフォームの出現をカウントするステップを含み、

前記収集したトレーニング文書中のキャピタライゼーションフォームの出現を該カウントするステップは、文書がユーザに関連付けられる度合いに基づいて出現カウントに重み付けするステップを含む、ことを特徴とする方法。

【請求項2】

前記制限事項は、前記トレーニング文書が前記ユーザによって書かれたものである、ことを必要とすることを特徴とする請求項1に記載の方法。

【請求項3】

前記ユーザが電子メールに返信した場合、前記電子メールは、前記ユーザによって書かれたものとみなされる、ことを特徴とする請求項2に記載の方法。

【請求項4】

前記ユーザが電子メールを転送した場合、前記電子メールは、前記ユーザによって書かれたものとみなされる、ことを特徴とする請求項2に記載の方法。

【請求項5】

前記制限事項は、前記トレーニング文書が前記ユーザのローカルマシン上に記憶されていることを必要とする、ことを特徴とする請求項1に記載の方法。

【請求項6】

前記制限事項は、前記トレーニング文書が前記ユーザに関連付けられたネットワーク上のディレクトリに記憶されていることを必要とする、ことを特徴とする請求項1に記載の

方法。

【請求項 7】

前記キャピタライゼーションフォームの出現をカウントするステップは、ワード対のうちの少なくとも 1 つのワードが大文字を含むワード対の出現をカウントするステップを含む、ことを特徴とする請求項1に記載の方法。

【請求項 8】

前記ワード対の出現をカウントするステップは、少なくとも 1 つの大文字を持つワード、および、前記ワードと少なくとも 1 つの大文字を持つ第 2 のワードとの間にある前置詞を含んだワード対の出現をカウントするステップを含む、ことを特徴とする請求項7に記載の方法。

【請求項 9】

前記ワード対の出現をカウントするステップは、少なくとも 1 つの大文字を持つ第 1 のワード、および、少なくとも 1 つの大文字を持つ隣接した第 2 のワードを含んだワード対の出現をカウントするステップを含む、ことを特徴とする請求項7に記載の方法。

【請求項 10】

前記収集したトレーニング文書を用いて前記キャピタライゼーションモデルを前記トレーニングするステップは、前記キャピタライゼーションフォームの出現のカウントを用いて、少なくとも 1 つのキャピタライゼーションフォームが前記キャピタライゼーションモデルに入れられないように該少なくとも 1 つのキャピタライゼーションフォームを除去するステップをさらに含む、ことを特徴とする請求項1に記載の方法。

【請求項 11】

前記収集したトレーニング文書を用いて前記キャピタライゼーションモデルを前記トレーニングするステップは、2 ワードを含むキャピタライゼーションフォーム中の大文字使用を含んだ各ワードが前記キャピタライゼーションモデル中の単一ワードと同じ大文字使用として現れる場合、前記キャピタライゼーションフォームが前記キャピタライゼーションモデルに入れられないように、少なくとも 1 つの前記キャピタライゼーションフォームを除去するステップをさらに含む、ことを特徴とする請求項1\_0に記載の方法。

【請求項 12】

前記キャピタライゼーションモデルを前記トレーニングするステップは、ワード対のためのキャピタライゼーションフォームをワード対リストに記憶し、単一ワードのためのキャピタライゼーションフォームを別個の単一ワードリストに記憶するステップを含む、ことを特徴とする請求項1に記載の方法。

【請求項 13】

前記キャピタライゼーションモデルを用いてテキスト中のワードを大文字化するステップをさらに含む、ことを特徴とする請求項1に記載の方法。

【請求項 14】

前記キャピタライゼーションモデルを用いてテキスト中のワードを前記大文字化するステップは、ワード対中のワードの 1 つを求めて単一ワードリストをサーチする前に、前記テキスト中のワード対に対するマッチングを求めてワード対リストをサーチするステップを含む、ことを特徴とする請求項1\_3に記載の方法。

【請求項 15】

文書が前記ユーザによって書かれている場合、前記出現カウントにより高い重みが適用される、ことを特徴とする請求項1に記載の方法。

【請求項 16】

前記キャピタライゼーションモデルがトレーニングされた後に、前記キャピタライゼーションモデルを更新するステップをさらに含む、ことを特徴とする請求項1に記載の方法。

【請求項 17】

テキストを自動的に大文字化するために、

特定のユーザに関連付けられた文書からトレーニングデータを獲得する手順と、

前記トレーニングデータ中の大文字ワードの出現をカウントし、かつ前記特定のユーザと、前記大文字ワードが現れる前記文書との関係に応じて、出現カウントに異なる重みを適用することによって、前記トレーニングデータを用いてキャピタライゼーションモデルをトレーニングする手順と、

前記キャピタライゼーションモデルを用いて、前記特定のユーザに関連付けられたテキストを自動的に大文字化する手順と、

をコンピュータに実行させるためのコンピュータ実行可能命令を記憶した、ことを特徴とするコンピュータ可読な記憶媒体。

**【請求項 1 8】**

前記文書からトレーニングデータを獲得する手順は、前記特定のユーザに関連付けられた電子メールからトレーニングデータを獲得する手順を含む、ことを特徴とする請求項1 7に記載のコンピュータ可読な記憶媒体。

**【請求項 1 9】**

前記電子メールからトレーニングデータを獲得する手順は、前記特定のユーザによって書かれた電子メールからデータを獲得する手順を含む、ことを特徴とする請求項1 8に記載のコンピュータ可読な記憶媒体。

**【請求項 2 0】**

前記電子メールからトレーニングデータを獲得する手順は、前記特定のユーザによって転送された電子メールからデータを獲得する手順を含む、ことを特徴とする請求項1 8に記載のコンピュータ可読な記憶媒体。

**【請求項 2 1】**

前記電子メールからトレーニングデータを獲得する手順は、前記特定のユーザによって返信された電子メールからデータを獲得する手順を含む、ことを特徴とする請求項1 8に記載のコンピュータ可読な記憶媒体。

**【請求項 2 2】**

前記文書からトレーニングデータを獲得する手順は、前記特定のユーザによって書かれた文書からデータを獲得する手順を含む、ことを特徴とする請求項1 7に記載のコンピュータ可読な記憶媒体。

**【請求項 2 3】**

前記文書からトレーニングデータを獲得する手順は、前記特定のユーザによって開かれた文書からデータを獲得する手順を含む、ことを特徴とする請求項1 7に記載のコンピュータ可読な記憶媒体。

**【請求項 2 4】**

前記トレーニングデータを用いてキャピタライゼーションモデルを前記トレーニングする手順は、大文字使用を伴う少なくとも1つのワードを含んだ大文字使用を伴うワードのシーケンスを求めて、前記トレーニングデータをサーチする手順を含む、ことを特徴とする請求項1 7に記載のコンピュータ可読な記憶媒体。

**【請求項 2 5】**

大文字使用を伴うワードのシーケンスを求めて、前記トレーニングデータをサーチする前記手順は、大文字使用を伴うワードだけを持つシーケンスを求めてサーチする手順を含む、ことを特徴とする請求項2 4に記載のコンピュータ可読な記憶媒体。

**【請求項 2 6】**

大文字使用を伴うワードのシーケンスを求めて、前記トレーニングデータをサーチする前記手順は、大文字使用を伴うワード、および前記トレーニングデータ中の大文字使用を伴う2ワード間にある前置詞だけを持つシーケンスを求めてサーチする手順を含む、ことを特徴とする請求項2 4に記載のコンピュータ可読な記憶媒体。

**【請求項 2 7】**

大文字使用を伴う単一ワードを求めてサーチする手順をさらにコンピュータに実行させるためのコンピュータ実行可能命令を記憶した、ことを特徴とする請求項2 4に記載のコンピュータ可読な記憶媒体。

**【請求項 2 8】**

大文字使用を伴う単一ワードの出現が、大文字使用を伴うワードのシーケンスの出現の一部を構成しない場合に限って、前記単一ワードの出現をカウントする手順をさらにコンピュータに実行させるためのコンピュータ実行可能命令を記憶した、ことを特徴とする請求項 2 7 に記載のコンピュータ可読な記憶媒体。