.Deb

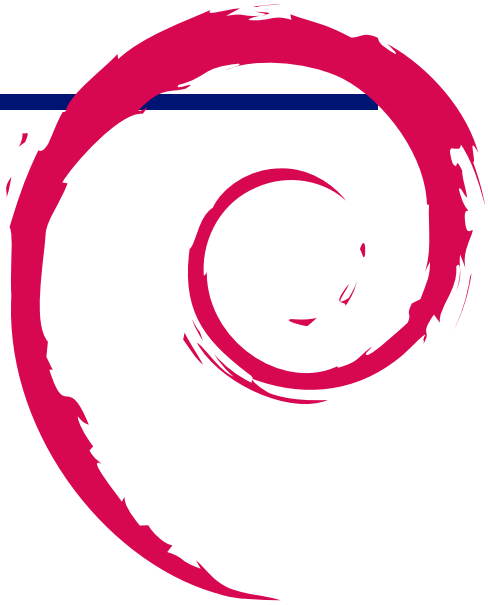
銀河系唯一のDebian専門誌

2015 年 5 月 23 日

特集：自然言語処理チームとパッケージ

デビアン勉強会

目次



1 事前課題

野島 貴英

今回の事前課題は以下です:

1. 本日、何の作業をやるかを宣言ください。
2. (オプション) どこで今回の勉強会の開催を知りましたか？
3. (オプション) 何について聞きたい／参加者と話をしたいですか？

この課題に対して提出いただいた内容は以下です。

1.1 野島

1. Q.hack time に何をしますか？
   A. Nook HD+ をそろそろ debian に。
2. (オプション)Q. 何について聞きたい／参加者と話をしたいですか？
   A. ソフトウェア自由を守り通す事に対して、何が出来るかについて。

1.2 wskoka

1. Q.hack time に何をしますか？
   A. MIPS Debian の移植
2. (オプション)Q. どこで今回の勉強会の開催を知りましたか？
   A. その他

1.3 koedoyoshida

1. Q.hack time に何をしますか？
   A. DDTSS,PyconJP 関連
2. (オプション)Q. どこで今回の勉強会の開催を知りましたか？
   A. 友達や知り合いから直接

DDTSS: http://ddtp.debian.net/ddtss/index.cgi/ja

1.4 dictoss

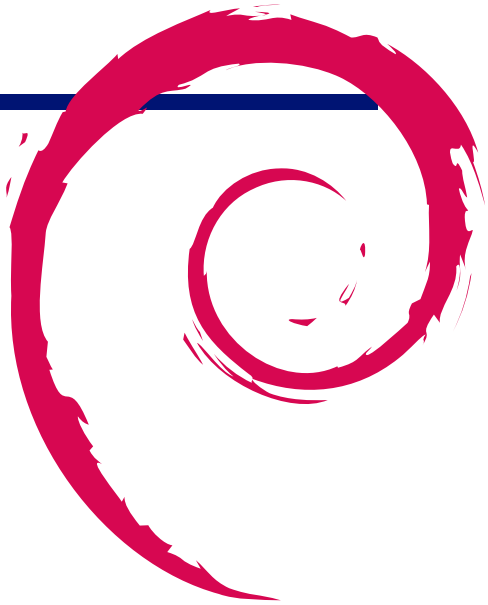
1. Q.hack time に何をしますか？
   A. kFreeBSD の IPsec 関連を調べる
2. (オプション)Q. どこで今回の勉強会の開催を知りましたか？
   A. Debian JP のメーリングリスト

1.5 yy_y_ja_jp

1. Q.hack time に何をしますか？
   A. DDTSS
2. (オプション)Q. どこで今回の勉強会の開催を知りましたか？
   A. その他
3. (オプション)Q. 何について聞きたい／参加者と話をしたいですか？
   A. DDTSS

1.6 myokoym

1. Q.hack time に何をしますか？
   A. Groonga のビルドをしながら、Debian で英語の形態素解析する方法を調べます。
2. (オプション)Q. どこで今回の勉強会の開催を知りましたか？
   A. Twitter (@tokyodebian)
3. (オプション)Q. 何について聞きたい／参加者と話をしたいですか？
   A. 英語の形態素解析のパッケージについて聞きたいです。



2 Debian Trivia Quiz

野島 貴英

Debian の昨今の話題についての Quiz です。

今回の出題範囲は `debian-devel-announce@lists.debian.org` や `debian-news@lists.debian.org` に投稿された内容などからです。

問題 1. Jessie が無事リリースされました。いつだったでしょうか？

☐ A 2015/4/11
☐ B 2015/4/18
☐ C 2015/4/25

問題 2. Jessie で初めて追加されたものではないものが混ざっています。どれ？

☐ A Debian Games Blend
☐ B OpenJDK
☐ C androidsdk-tools

問題 3. Debian GNU/Hurd 2015 も 2015/4/30 にリリースされました。心臓部の GNU Mach のバージョンはいくつ？

☐ A 1.6
☐ B 1.5
☐ C 1.4

問題 4. 2015 年の GSoC に採択された、Debian MIPS ports についての開発内容は次のどれ？

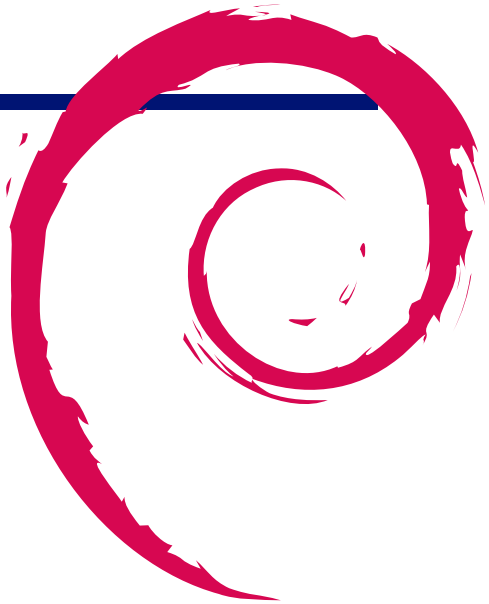
☐ A 多数のビルド出来ないパッケージを、ちゃんとビルドできるようにする。
☐ B 新しい MIPS CPU への対応
☐ C 新しい MIPS CPU 搭載製品への対応

問題 5. 遂に http.debian.net が debian.org のインフラに移動となりました。新しい URL はどれ？

☐ A http://http.debian.org/debian
☐ B http://httpredir.debian.org/debian
☐ C http://www.debian.org/

問題 6. debian への libdvdcss/ZFS パッケージ搭載の件について、2015/5 の Bit From DPL で報告された状況は以下のどれ?

☐ A DPL がいろいろ議論して回っている状況
☐ B 搭載日確定した
☐ C 搭載を諦めた



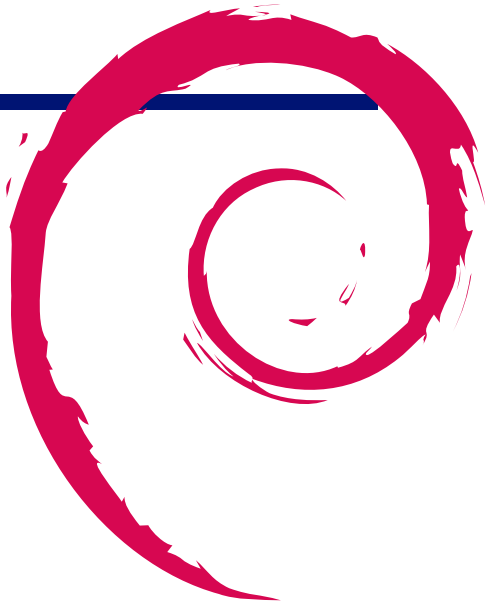
3　最近の Debian 関連のミーティング報告

野島 貴英

3.1　第 125 回東京エリア Debian 勉強会

- 場所はスクウェア・エニックスさんのセミナルームをお借りしての開催でした。
- 参加者は 13 名でした。
- セミナ内容はまえだこうへいさんによる「.deb から Python パッケージへの変遷」でした。
- 残りの時間で hack time を行い、成果発表をしました。
- 宴会の代わりに、「祥龍房 新宿イーストサイドスクエア店」で夕食会をやりました。

セミナはまえださんにより、Python パッケージのローカル Debian CI に関する発表と、Python に関する Debian 公式パッケージの考察が行われました。大統一 Debian 勉強会から、毎年発表のはや 3 年越しのシリーズものとなります。昨今の言語関係の流行りや事情から、言語のパッケージを Debian のパッケージ化を検討した方が良い場合と、あまり向かない場合がある状況とのことです。会場に Haskel の詳しい人、Perl に詳しい人がおり、様々に議論がなされました。



4 自然言語処理チーム (pkg-nlp-ja) とパッケージ

野首 貴嗣

4.1 NOTICE

私は専門家ではありません。

4.2 pkg-nlp-ja

Debian における自然言語処理パッケージの管理 `https://alioth.debian.org/projects/pkg-nlp-ja/`

4.3 管理パッケージ

- ChaSen
  - dirts (Double Array 実装)
- 辞書
  - ipadic
  - naist-jdic
  - unidic

4.4 その他の NLP パッケージ

- 解析
  - MeCab
  - Juman
- 変換
  - Anthy (pkg-anthy)
  - KAKASI
  - mozc
    - 変換については pkg-ime で管理

4.5 何ができるか

形態素解析

- 形態素 (単語) を調べる
- 単語の属性情報を得る
  - 品詞・読み等
    - 辞書に依存

ChaSen, MeCab, Juman ref:`http://www.phontron.com/nlptools.php?lang=ja`

4.6 応用例

- 検索エンジン
  - 転地インデックスの作成
  - Groonga, Namazu 等
- 音声合成 (Text-to-Speech)
  - open-jtalk

  n
- 分かち書き
  - word2vec の下処理
- かな漢字変換

4.7 動作原理

- 解析対象を確からしい単位で分割
  - 例: 「東京都府中市」
    - x 「東」「京都」「府」「中」「市」
    - o 「東京」「都」「府中」「市」
  - 単純なルールでは処理できない

4.8 解析の手がかり

- マッチする単語の長さ
  - 最長一致 (KAKASI)

  ref: KAKASI の実装と課題
- 品詞情報の利用
  - 前後に来やすい品詞のつながり
  - 単語のつながり
  - 単語の頻度/スコア
- 確率
  - 出現率、連接確率

4.9 アルゴリズム

- コスト最小法
  - ChaSen
- CRF (条件付き確率場)
  - MeCab

ref: 日本語形態素解析入門 (pdf)

4.10 辞書探索

Common Prefix Search

- 同じ接頭語を持つデータを高速に検索

アルゴリズム

- トライ
  - パトリシアトライ
- Double Array
  - darts パッケージ

4.11 トライ (Trie)

文字単位の木構造

```
漢-+->字-+->化
   |     |
   |     +->語
   +->音
```

ref: http://ja.wikipedia.org/wiki/基数木

4.12 Double Array

トライ構造を 2 つの配列で表現

- BASE 配列 (子ノード番号へのオフセット)
- CHECK 配列 (親ノード番号)
  - 単純な状態遷移表だと疎な配列になる
  - 構築に計算が必要
    - 配列の空き要素を見つける必要がある
    - 動的な更新には向かない

ref: ダブル配列の実装方法

4.13 辞書

- KAKASIDIC(kanwadic)
  - SKKDIC ベース
- ipadic
  - ChaSen, MeCab
- naist-jdic
  - ChaSen, MeCab
- jumandic
  - MeCab, JUMAN

4.14 単語の追加: KAKASI

辞書ファイルの作成 (EUC-JP)

```
# よみ [空白] 漢字
けいさんしょう 経産省
けいざいさんぎょうしょう 経済産業省
```

kakasi コマンドの引数に辞書のパスを追加

```
$ echo '日本の経産省' | kakasi -w -iutf8 -outf8 ./extdic
日本 の 経産省
$ echo '日本の経産省' | kakasi -w -iutf8 -outf8
日本 の 経 産 省
```

4.15 単語の追加: ChaSen

- ipadic のソースを取得
- 追加したい単語の意味に近いものを探す
- そのエントリーをコピーし、スコアはそのままで単語を修正する
- ./configure && make

注: スコアを計算するツールが付属していない

4.16 単語の追加: MeCab

1. ChaSen と同じ方法
   - 活用のある品詞はすべて開く必要がある
2. mecab-cost-train で正確なコストを算出
   - 元となるコーパスを用意する必要がある
   - cf: mecab-ipadic-neologd
   - MeCab の IPA 辞書を再学習させてみる
   - MeCab の辞書をカスタマイズする

4.17 参考

自然言語処理ツール

`http://www.phontron.com/nlptools.php?lang=ja`

日本語で読める自然言語処理のチュートリアルスライドまとめ

`http://blog.unnono.net/2015/04/nlp-tutorial.html`

Debian 勉強会資料

2015 年 5 月 23 日　　初版第 1 刷発行

東京エリア Debian 勉強会 （編集・印刷・発行）